UNIVERSITY OF PLYMOUTH

# AUTHOR VERIFICATION OF ELECTRONIC MESSAGING SYSTEMS

by

ABDULAZIZ ALTAMIMI

A thesis submitted to the University of Plymouth

in partial fulfilment for the degree of

DOCTOR OF PHILOSOPHY

School of Engineering, Computing and Mathematics

July 2020

# Copyright Statement

# Acknowledgements

First and foremost, all praise and gratitude are due to Allah Almighty, the All Merciful, for helping me in tackling all the challenges faced throughout this PhD.

I am deeply indebted and most of my sincere thanks and appreciation go to my beloved parents, for their considerable help and support, kindness, and prayers for my study, and I ask Allah to reward them with the best.

I also owe many thanks to my beautiful love, my wife Aziza, and my children, Tamim, Obaid, Layan and Rawan, for their patience, endless support, and incredible care in assisting me throughout this endeavour. They all stood alongside me and provided me with an abundance of love and support, even when spending hours, nights and holidays without me. I really appreciate your limitless support and help throughout this PhD journey. I am ever grateful.

Of course, I would like to extend my most sincere thanks and heartfelt appreciation to my Director of Studies, Professor Nathan Clarke, for his patience, motivation, wisdom, help and a sympathetic ear. I am grateful to have a supervisor who cared so much about my work, and who responded to my questions and queries so promptly. I would also like to extend my most sincere thanks and my deepest appreciation to my supervision team, Professor Steven Furnell and Doctor Fudong Li, for their guidance, support and help. Their experience and professionalism in various aspects, such as their critical thinking and publications, have been invaluable throughout my PhD journey and without their valuable comments and advice, I would not be able to make this a success, so thank you.

I would also like to express my thanks to my research colleagues at the Centre for Security, Communication and Network Research, Abdulrahman Alruban, and

Saud Alotaibi, who have been my motivation and inspiration and with whom I have held interesting discussions during this PhD journey.

I would like to take this opportunity to thank all my colleagues at the Ministry of Higher Education in the Kingdom of Saudi Arabia for allowing me to take this great opportunity to complete my PhD degree and granting me a scholarship and sponsoring my PhD studies.

# Author's Declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award without prior agreement of the Doctoral College Quality Sub-Committee.

Work submitted for this research degree at the University of Plymouth has not formed part of any other degree either at the University of Plymouth or at another establishment.

This study was financed with the aid of a studentship from the Royal Embassy of the Kingdom of Saudi Arabia.

Relevant scientific seminars and conferences were attended at which work was often presented and several papers were published and prepared for published.

Word count of main body of thesis: 75,122 words

**List of Publications:**

Altamimi, A., Clarke, N., Furnell, S., & Li, F. (2019, November). Multi-platform authorship verification. In Proceedings of the Third Central European Cybersecurity Conference (pp. 1-7).

DOI: https://doi.org/10.1145/3360664.3360677

Signed: Abdulaziz Altamimi

Date: 22 July 2020

# Abstract


## AUTHOR VERIFICATION OF ELECTRONIC MESSAGING SYSTEMS

## ABDULAZIZ ALTAMIMI

Messaging systems have become a hugely popular new paradigm for sending and delivering text messages; however, online messaging platforms have also become an ideal place for criminals due to their anonymity, ease of use and low cost. Therefore, the ability to verify the identity of individuals involved in criminal activity is becoming increasingly important. This field of study is known as authorship verification. The majority of research in this area has focused on traditional authorship problems that deal with single-domain datasets and large bodies of text. Few research studies have sought to explore multi-platform author verification as a possible solution to problems around forensics and security. Therefore, this research has investigated the ability to identify individuals on messaging systems, and has applied this to the modern messaging platforms of Email, Twitter, Facebook and Text messages, using different single-domain datasets for population-based and user-based verification approaches. This research has also explored unifying author features across platforms and the relationships that exist within linguistics cross-domain. Through a novel technique of cross-domain research using real scenarios, the domain incompatibilities of profiles from different distributions has been assessed, based on real-life corpora using data from 50 authors who use each of the aforementioned domains. A large sample size was used, as the total number of samples in each corpora was 13,617; 106,359; 4,539 and 6,540 for Twitter, Text message, Facebook and Email respectively. In addition, the volume of information needed to provide a reliable way of determining an author's identity has been explored, along with the level of confidence in an author verification decision.

The results show that the use of linguistics is likely be similar between platforms, on average, for a population-based approach. The best corpus experimental result achieved a low EER of 7.97% for Text messages, showing the usefulness of single-domain platforms where the use of linguistics is likely be similar, such as Text messages and Emails. For the user-based approach, there is very little evidence of a strong correlation of stylometry between platforms, meaning that

users communicate quite differently with different sets of stylometry on individual platforms. It has been shown that linguistic features on some individual platforms have features in common with other platforms, and lexical features play a crucial role in the similarities between users' modern platforms. In addition, for the volume of information needed to provide a reliable determination of an author's identity, on average, the best performance was for Text messages, with an EER of 7.6% if the number of words was more than nine; followed by Email with an EER of 14.9% if the number of words was between 25 to 60; then, Twitter tweets, with an EER of 22.5% if the number of words was less than ten. Finally, the Facebook platform with an EER of 31.9% if the number of words was over 11.

Therefore, this research shows that the ability to identify individuals on messaging platforms may provide a viable solution to problems around forensics and security, and help against a range of criminal activities, such as sending spam texts, grooming children, and encouraging violence and terrorism.

This research investigates the ability to identify individuals on messaging systems, and how this can be applied to modern messaging platforms. This is becoming increasingly important for a number of reasons, for example, a suspect may have an ordinary Facebook profile with which he/she communicates with friends, yet may perform criminal activities on another different platform such as Twitter; alternatively, it is also possible for a criminal or other user to send a message on behalf of the real user. A suspicious message from an individual's platform can be viewed by families, friends, or by anyone on the messaging platforms that are hosted by the real author's messaging systems. In order to investigate authors using messaging platforms, a method is required to reduce such security threats. Therefore, this research is an attempt to investigate the ability to identify individuals on electronic multi messaging systems.

# Table of Contents

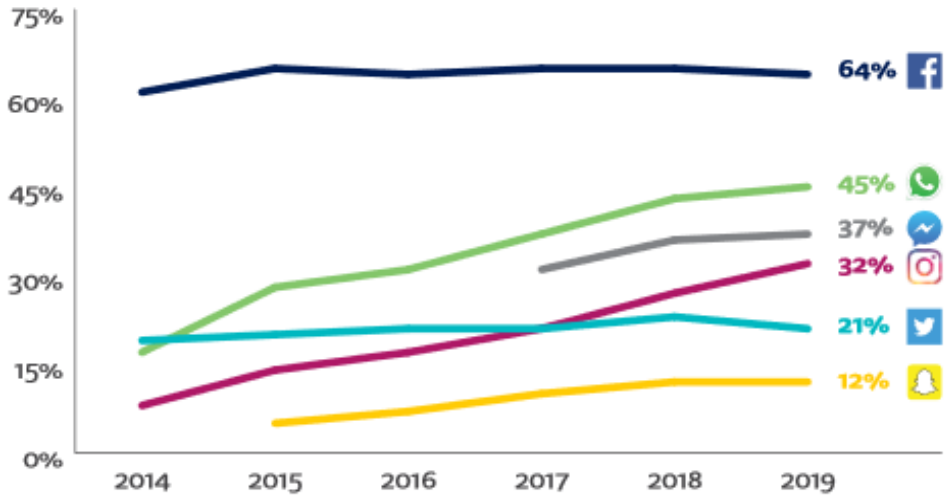
# List of Figures

# List of Tables

# Chapter One: Author Verification of Multiple Messaging Platforms

## 1.1 Introduction and Overview

Around 500 million tweets are sent, and 4.3 billion Facebook messages are posted, every day; in addition, more than 200 million emails are sent and approximately two million new blog posts are created daily, and around 15 billion texts are sent every minute around the globe (Schultz, 2019; Smith, 2019). Indeed, research has shown that it is popular (typically for someone in their 20s) to utilise multiple messaging systems (Almishari et al., 2015). For example, a study by Chung (2014) reports that 64% of Facebook users also had accounts on Myspace, and LinkedIn shared 42% of its members with Facebook and 32% with Myspace. Figure 0-1 illustrates the rapid growth in several popular social network and messaging systems from twelve countries during 2014-2019, demonstrating an increasing trend of usage across these messaging systems. For example, the number of active users of WhatsApp increased from 15% of the entire population in 2014 to 45% in 2019.




Source: (Reuters Institute Digital News Report, 2019)

Figure 0-1: Annually active users on selected social networks

However, despite the popularity of messaging systems, they are often found to be the source and target of criminal activities. Messaging systems have become an ideal place for criminals due to their characteristics such as anonymity, ease of use and low cost (Chen , 2018; Cai et al.,2020). This leads to a variety of direct and indirect criminal activities, such as sending spam texts to gain personal information (Stringhini., 2010), grooming children, kidnap, murder, terrorism and violence (Page et al., 2014; Weir, 2011). For example, an analysis of the London riots in 2014 shows that Twitter was used to provide key command and control functionality and services for criminals (Ball et al., 2011; Tonkin et al., 2012). This is the first documented example in the UK of a messaging system being used to facilitate widespread criminal activities. Also, the problem of online trolling, which has become a concern in the UK, has increased with the rapid spread of social networks (Roberts et al., 2017). Globally, along with the events that have taken place in the Middle East such as the so-called Arab Spring, Twitter Uprising or Facebook Revolution that occured in Egypt, Syria, Libya, Yemen and Tunisia in 2011, which all relied on social messaging systems, social messaging networks have become part of daily (sometimes violent) activities, and have opened up a new era of informal messaging communication (Shearlaw, 2016; Salim, 2012; Ward, 2018).

The Daesh extremist terrorist group, or so called Islamic State (IS), under the pretext of the religion of Islam, has used Twitter and Facebook as well as mobile applications such as Telegram, WhatsApp, Wickr and SureSpot, to broadcast its threating text messages to Saudi Arabia and others countries and to send invitations via text messages (SMS) to Saudi teenagers to invite them to conflict zones (Bodine-Baron, 2016; Jawhar, 2016). In France, hundreds of thousands of

protesters started to gather in one place using Facebook and Twitter to plan to damage and destroy property and government assets, with reporters describing the Yellow Vest movement as a feedback loop that started on Facebook and generated violent protests in the real world (Newton., 2018; Peltier et al., 2018). The Irish Republican Army (IRA) terrorist group use regular the internet and social media to organise its activities and to spread their propaganda goals (Tsesis., 2017). Messaging platforms such as Facebook, Twitter, WhatsApp, Instagram, Text message and Email, are facilitating the spread of information more quickly and easily, and can be channels to deliver an evil message in a clear and straightforward way to the intended party and their target.

Whilst this thesis focuses on the negative aspect of social media in regard to finding the sources of criminal activity and individuals behaving in a non-ethical manner online, there are also many benefits that can be gained from social networks. For example, According to Joo (2017) consider the pros and cons of social media and point out how it can be used to combat loneliness and bring families who live together. On the other hand, social networks may harm relations, as it is possible for family members to sit in the same room whilst not communicating with each other due to focusing on social media, typically via mobile phones (Joo et al., 2017). While social networks allow a wide range of social, political and environmental views to be shared, there is also the danger of people existing inside an “echo chamber” where they share their views only with like-minded others (Harris et al., 2015). On the other hand, the internet allows people to connect with like-minded others, which may be useful for their emotional wellbeing.

Social networks can extend a person’s friendship circle and allow them to share more effectively, both from a social and business perspective. However, it is

important to consider the safety of such interactions, due to new phenomena such as cat-fishing, where the other person takes on a fake personal (Reichart et al., 2017). This means that it is important to attempt to check the identity of unknown persons online, especially if personal information is being shared or plans are being made to meet in person. On the other hand, social networks can provide support that is not readily available in the real world, especially for minority groups in society, such as people with the same disability connecting with others and sharing their experiences, or refugee communities reaching out to each other and sharing advice (Hanley et al., 2018). Therefore, social networks have both positive and negative consequences, and it is important to take advantage of the aforementioned benefits while ensuring the safety of users and working towards discovering those who use social networks in an unethical or criminal manner. Thus, this research in the current study should help to both discover individuals engaging in illegal activity, and perform as a deterrent to those planning to do that in the future. The aim is to increase the safety of social networks so the people can take advantage of its benefits.

A need exists, therefore, to be able to identify the ownership of messages shared on these electronic systems. Unfortunately, relying on just the account details to simply verify the author of that account could be misleading because messaging platforms typically do not enforce identity checking, thereby enabling the creation of fake accounts or accounts which are not easily traced back to an individual (Nirkhi et al., 2012; Maheswaran et al., 2010). Authorship verification is, however, an approach that provides the ability to determine the authenticity of the author through an examination of the message. Author verification is not a new research area - it has been used in the past to verify and identify authors from many aspects and in a range of studies (e.g. Brocardo et al., 2017; Saevanee et al., 2011; Li et

al., 2014; Abbasi et al., 2008; Koppel et al., 2013; Iqbal et al., 2010; Ragel et al., 2013; Nirkhi et al., 2012; Silva et al., 2011; Zafarani et al., 2013).

## 1.2 Research Contributions

- Investigated the performance of features by verifying the authorship of a given text message across different messaging platforms. The impact the feature vector has on performance has also been explored, as well as the performance of authorship verification across a number of common messaging platforms. This includes an exploration of the classification performance of population and individual based verification approaches.
- Explored the viability and the ability to use feature vectors derived from one or more messaging platforms on other messaging platforms. The aim is to gain an understanding of differences in linguistic characteristics by examining their performance and combining linguistic feature verification, to find the top discriminatory features for a user of a variety of platforms under a particular umbrella mechanism - the features have been unified and integrated with each other on different platforms to verify the particular user. After that comes the portability of the authorship of different text messages (portability linguistic feature verification) to test a text sample against another sample from a different platform, called cross-domain datasets, in order to investigate the process of common features of a user profile across platforms. By verifying different sets of features for a given author's sample across various messaging systems, the common features can be identified.
- Examined and analysed the message length required to enable reliable author verification decisions.

## 1.3 Research Aim and Objectives

The aim of this research is to explore the application of authorship verification to electronic messaging systems[1]. It consists of two main objectives: The first main objective is to explore authorship verification within these electronic messaging platforms by comparing relative performance across messaging systems, and the degree to which author verification is possible when using different single-domain datasets for population-based and user-based verification approaches. The second main objective is to explore the unification and portability of author features across platforms in order to understand what relationship, if any, might exist within linguistics on multi-platforms. In addition, the investigation has also sought to identify the minimum amount of text required whilst still ensuring a reliable performance for each platform. In order to achieve this aim, the following research objectives have been set:

- Conduct an exhaustive literature review of the existing research in the domain of identification and verification techniques, focusing specifically on short-text approaches, to understand and evaluate the current state of the art author verification and identification techniques on different platforms.
- Conduct a series of experiments aimed at investigating authorship verification in a platform independent manner, including assessing how well author verification performs on individual platforms, and exploring feature vector composition, as well as the impact of classification on performance.

[1] Electronic messaging systems is defined as a digital system that allows people to send and receive messages (e.g. SMS text messages, Facebook posts, WhatsApp messages, Twitter tweets, and Email correspondence).

- Conduct a series of experiments aimed at exploring the portability of feature vectors in three ways: looking at the ability to use a profile against other platforms, the creation of unified feature vectors, and the portability of features.
- Examine and analyse the minimum set of information that would be required to provide reliable verification of an author. This would measure and characterise the limitations with respect to message length and composition to provide reliable author verification decisions.

## 1.4 Thesis Structure

This thesis is organised into seven chapters to address the above-mentioned objectives, commencing with Chapter one, which introduces the research problem and outlines the overall research aim and objectives and the structure of this thesis.

Chapter two sets out the background information about biometric systems, characteristic components, requirements, techniques and performance measures, with a specific focus on the core background knowledge on author verification.

Chapter three presents a comprehensive literature review in the domain of author verification of text, including the types of stylometric features notable in long text and as well as short text, along with author verification for different domains. This is accompanied by a exploring the size and length of the messages required to facilitate the process of verification. The chapter concludes with a discussion that identifies the gap that exists in the literature by highlighting the need for new security mechanisms to assist investigators and improve author verification on electronic messaging systems.

Chapter four contains the research methodology applied, and includes the methods used to answer the two main research questions; the steps taken in the research, and the methods used to conduct the experiments. It also shows the methods used to collect real data from the core messaging systems and online social networking platforms. The goal was to collect as many text messages as possible, and the research has targetted users of at least two platforms, with a minimum of 20 messages available on each corpus. In addition, this chapter discusses feature selection and extraction, and the methodology and data collection process.

Chapter five presents the first experiment on verifying the authorship of a given text message in order to identify what stylometric features are apparent on all platforms and for individual authors. This includes two types of experiments for verifying the authorship of a given text message (Twitter, Text message, Facebook, Email). Population and user feature-based verification approaches have also been examined. A series of experiments were also conducted with different settings of different classifiers to investigate the portability of user profiles across messaging systems.

Chapter six presents the second experiment on unified author verification profiles approach, and contains an extensive and comprehensive investigation to unify and give an overall picture of user profiles on the different platforms. The main objective is to create common stylometric features that can help to find a user's common features on various platforms. It contains three sets of investigations: the first experiment aims to unify the user profile on all platforms and verify authors. The second experiment investigates the portability of user profiles and tests these against each other on the different types of platform. The third experiment presents the minimum set of information that would be required to

provide reliable verification of an author. This has facilitating measuring and characterising the limitations with respect to message length and composition to provide reliable author verification decisions.

Chapter seven is the final chapter, and it summarises the conclusions arising from the research and underlines the key contributions, achievements and limitations. It also contains a discussion of potential areas for future research.

A number of appendices are included at the end of this thesis in support of the main discussion, containing experimental ethical approval, consent forms, stylometric features list and programming scripts, and a number of published papers arising from the research programme.

# 2. Chapter Two: Background on Biometric Systems and Author Verification

Author verification is a form of biometric system that discovers an individual's identity based on their writing style. Writing styles vary (to a certain degree) from one person to another, and such variations can be utilised to verify the authorship of written text. In order to be able to understand the nature, characteristics and performance of author verification, it is prudent to investigate how biometric systems work and operate. Therefore, this chapter will proceed by first introducing the background knowledge on biometric systems and modalities, before proceeding to describe the fundamental background of author verification.

## 2.1 Biometric Systems

Biometrics is an approach that has been utilised to automatically authenticate and identify individuals based on their unique characteristics to access many digital systems, such as the use of facial recognition on smartphones (Clarke and Furnell, 2007); the use of iris recognition at passport control (Gorodnichy, 2014), and the fingerprint recognition by law enforcement (Sankaran et al., 2017). In this electronic information technology world in which people work and live, there is an increasing need to employ biometrics to secure users' assets and reduce security threats (Ali & Awad, 2018).

Biometric modalities can be divided into two types: physiological and behavioural biometrics (Jain et al., 2004). Physiological biometric recognition is built on the basis of the human body, for example, the shape of a face, ear or hand. Alternatively, behavioural biometrics are built on the basis of a person's behaviour, for example, the way they perform a certain task, and the way in which they write, sign, speak and walk.

It is important to indicate that there are two functional modes of biometrics systems, which are verification and identification (Furnell & Clarke, 2005), as described below:

- Verification (Does this identity belong to you?): the system attempts to authenticate that a claimed identity is matched with a profile on the database. In this system, the user needs to claim an identity and the system will conduct a one-to-one comparison to determine whether the claimed identity matches with the user template.
- Identification (Whose identity is this?): the system will identify the user by searching a matched identity from all of the biometric reference templates stored on a database. The system conducts one-to-many comparisons to determine the identity of the template that matches.

Verification is widely used for authenticating individuals at point of entry systems, and to verify individuals' identity; while identification is extensively employed in forensics and law enforcement activities to identify uncooperative suspects or provide covert identity checking. For example, fingerprints have been used in many law enforcement investigations (Sajjad et al., 2020; Jain et al., 2010); facial identification has been used to monitor terrorist watch lists (Introna et al., 2010), and speaker recognition has been used for identifying criminals on the telephone (Singh, 2018). Author attribution has also been used on occasion to identify or verify the identity of individuals from Email messages, and this has been used by the police in the UK (Wright, 2014; Ensor, 2013).

## 2.2 Biometric Characteristics

Biometric applications have become the foundation of verification and identification for many applications (Akhtar et al., 2017). A number of factors

should be considered when a specific biometric is being used within a specific application. These factors or characteristics have an important effect on the biometric system security, and influence matching decisions, level of uniqueness and performance. The characteristics are (Jain et al., 2007):

- Universality: Every person using the technique should have the main characteristic, for instance, people need to have fingers for the fingerprint method to be used. On the other hand, there are many biometric techniques where a subset of the population do not have the biometric traits in order to provide the sample, such as people with missing or damaged fingerprints (Sharma et al., 2013).
- Uniqueness or distinctiveness: individual modalities have varying degrees of discrimination or uniqueness, and the modalities should be sufficiently different for an individual from amongst the population such that they can be discriminated from one another. For example, physiological-based techniques such as fingerprint and iris recognition tend to be far more discriminatory than their behavioural counterparts such as keystroke analysis or signature recognition (Kour, 2011).
- Permanence: This refers to the ability of the biometric characteristic to remain constant over time. For example, the fingerprint is a physical biometric whose features remain stable over the time, whilst some behavioural biometric techniques such as signature or keystroke analysis are subject to change over time.
- Collectability or measurability: this means the ability to reliably collect a sample of sufficient quality that can be used on a reliable basis.
- Performance: this indicates how well a recognition system performs in identifying individuals.

- Acceptability: this refers to what extent people are willing to accept it or avoid using it. For example, in the case of iris recognition, some people may perceive that the capturing of the sample might be harmful due to the exposure to infrared or a laser in the scanning process (Obaidat et al., 2019) .
- Circumvention: this means to what the degree a trait is vulnerable to being forged. For instance, fingerprints can be forged through the use of plastic and latex fingerprints (Meghanathan , 2018).

It is expected that every biometric technique will exhibit each characteristic to some degree; however, in practice, the degree to which a particular technique meets them can vary considerably (Furnell et al., 2005). Moreover, human behaviour can be unstable over time because of certain factors such as age, health, mood and social situation, which will also impact the characteristics of the biometric analysis and the sensitivity/acceptance of the basic biometric method (Bolle et al., 2013). Therefore, care is required in terms of selecting the most suitable technique that has an appropriate mix of the characteristics required for the particular application within which it will be deployed.

## 2.3 Components of Biometric Systems

There are four basic components for every biometric system which are shown in Figure 2-1 below. According to Clarke (2011), these components are:

- Sample Capturing: This component is the first stage of the biometric system that captures the biometric sample from a user by using a sensor device such as a fingerprint scanner for individual fingers.

- Feature Extraction: The extraction process extracts a set of unique biometric features from the captured sample to create a template, and subsequently stores it on the database.
- Matching or Classification: In the classification phase, the system will compare the newly captured sample with the stored reference template(s), with the output of this component showing the degree of similarity between the two samples.
- Decision: This is the final phase of a biometric system. Its purpose is to decide whether the resulting score is sufficient to meet the level of security and usability of the system. This is achieved through setting a threshold which a similar score must meet or exceed to be accepted.

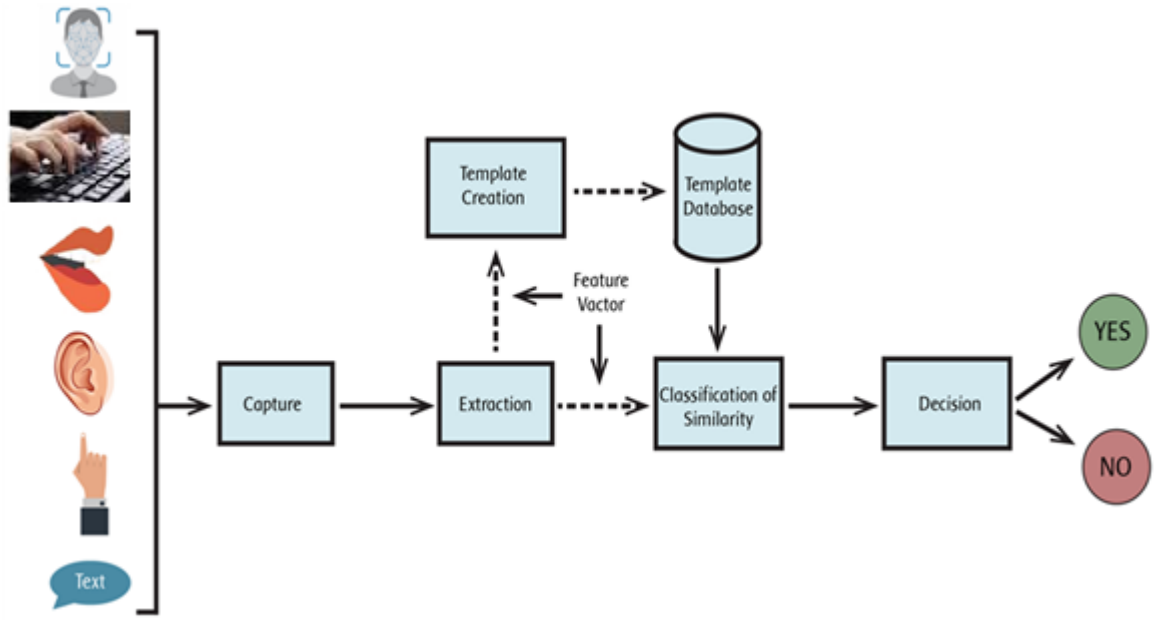



Figure 2-1: The biometric system components

These biometric components are supported by two processes, which are enrolment and verification/identification. In the enrolment process, a user can register to a biometric system, and the system captures the biometric sample from the user to create the reference template. Therefore, it is important at this stage of the process that the sample is of sufficient quality and from a specific legitimate

individual, as this sample needs to be reliable in future verification (Jain et al., 2004).

After the enrolment process has been completed, the biometric system is ready to perform two different modes: The mode of authentication/verification verifies a claimed identity, and the mode of identification determines the identity (Bolle, 2013). The enrolment, verification and identification processes are shown in Figure 2-2.

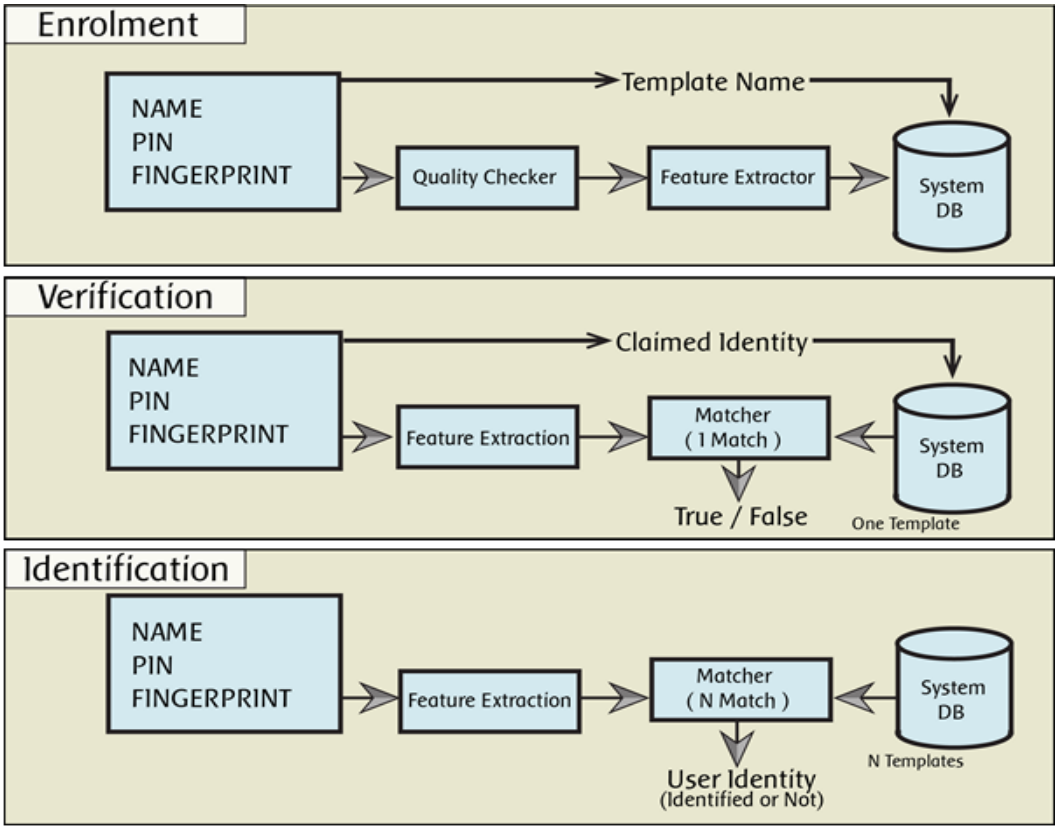



Figure 2-2: Enrolment, verification and identification processes in a biometric system

For verification, the system matches the collected biometric of a user that a user claims to be the claimed identity that is before stored in the template of the database system. If the match is successful, the person can access to the system; otherwise, access is denied. The process of verification is referred to as one to one (1:1) matching. For instance, when a user wants to login to a computer system by using their username and fingerprint, he or she will enter a username,

and then provide a biometric sample by scanning his or her finger. The new captured sample of the fingerprint will be compared with the sample template which was previously stored on the database based on the user name given. If they match with each other to a sufficient degree, the user is given access, and if not, the user will be rejected.

For identification, the process is similar to verification, but the difference occurs in the matching process, where the user does not claim an identity but the system matches the sample against all enrolled users to identify whether there is any match present or not. The process of identification is referred to as one to many (1:N) matching. A personal identification number (PIN), login name, smart card, or other identifier, is required in order to establish an individual's identity for the system to conduct a one-to-many comparison (if it fails, the subject is not enrolled on the system's database). For example, when the criminal investigator is investigating someone and needs to find out the previous activities of this criminal that have been stored on a criminal database, and to access this individual's activities through his or her face, this face sample of the criminal is compared with all stored criminals' faces to decide whether there is a match or not.

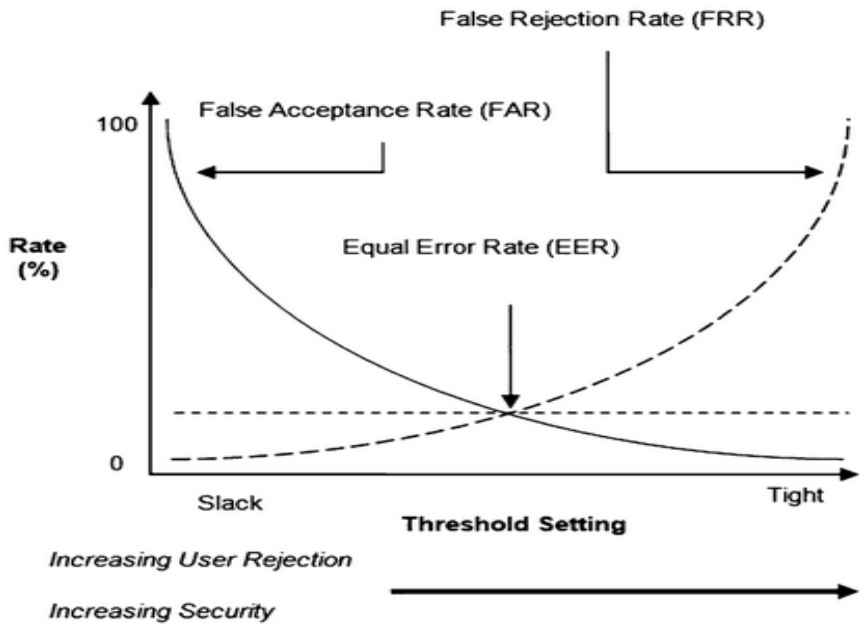
In general, all biometric systems contain these fundamental components; however, in practice they are often accompanied by additional components that aid in the overall process. For example, it is typical to have pre-processing stages that allow for noise reduction, segmentation and the normalisation of signal inputs (Nelufule, 2014). The nature of these, and whether they are required, is modality dependant.

## 2.4 Biometric System Performance

Evaluating the performance of a biometric system is an important task to measure and compute the error rate that such a system has. There are two fundamental evaluations that can be used to assess the operation of biometric systems: First, in verification mode, the two main error rates can be used, but it is not restricted to them; these are: the False Acceptance Rate (FAR), and False Rejection Rate (FRR).

- False Acceptance Rate (FAR) describes the probability that an imposter is falsely accepted by the system.
- False Rejection Rate (FRR) describes the rate of the system rejection to the authorised user who attempts to access the system.

The relation between FAR and FRR is a mutually exclusive relationship, which results in a challenge to reduce both to zero, because when the values of FRR decline, the values of FAR increase. In addition, there is a specific point when both the FAR and FRR overlap, which is the Equal Error Rate (EER), as shown in Figure 2-3. The EER is often utlised as a means of comparing the peformance of different biometric systems.




(Source: Clarke, 2011)

Figure 2-3: Biometrics performance metric factors

On the other hand, the FAR/FRR metrics are more inclusive in that they also involve the Failure to Enrol (FTE) and Failure to Capture (FTC) rates.

**FTE:** shows the rate of unsuccessful biometric registrations where individuals are unable to create an initial template (Furnell & Clarke, 2005).

**FTA:** shows the rate of a biometric sensor's device failure to acquire/capture a biometric sample and locate it on the templates database (Clarke, 2011; Jain et al., 2008).

Both errors may be caused by a number of reasons, including but not limited to: missing the main related trait (e.g. missing a finger or hand completely); poor quality of the biometric sample that could be attributed to sensor insufficiency (e.g. those caused by wear and tear); user mistakes (e.g. wrong posture for facial recognition); or inconsistent measurement pattern (e.g. changing on the way the user types on a keyboard); environmental effects (e.g. poor lighting or noise).

Furthermore, there are other error rates that can be utilised to evaluate biometric systems, for instance, True Acceptance Rate (TAR) illustrates the ratio at which the system correctly verifies the claimed user, and true rejection rate (TRR) illustrates the ratio at which the system rejects a false claim.

The second approach to performance evaluation is used to measure the operation of the biometric system in the identification mode:

- True positive identification rate (TPIR) illustrates the ratio of identification transactions by registered users who are registered on the biometric system in which the user's correct identifier is among the returned matches
- False Negative identification rate (FNIR) presents the ratio of identification transactions by registered users who are registered on the biometric system in which the user's correct identifier is not among the returned matches [FNIR = 1 – TPIR].
- False positive identification rate (FPIR) is the ratio of identification transactions by users not registered on the biometric system, in which an identifier is returned.

## 2.5 Behavioural Biometric Techniques

As previously highlighted, there are two types of biometric approaches: physiological and behavioural. Physiological biometrics are related to some unique physical characteristics of the human body, such as the fingers, face, eyes and hands, while behavioural biometrics are related to certain behaviours of the person, such as their signature, speech, gait and writing. Physiological biometrics have been extensively researched and are currently used in a wide variety of applications because of the high degree of accuracy that they can achieve (Dargan et al., 2020); although biometric behaviours are typically more

acceptable and easier to collect. Author verification is a form of behavioural approach; therefore, this section will focus on behavioural biometric techniques, and the following section illustrates, in more detail, the linguistic profiling techniques used for author verification and identification.

Behavioural biometrics can be applied to many different applications for verification and identification, and can offer adequate discrimination between users, thus leading to the recognition of users.

Based upon the seven characteristics that are required for a biometric system, as illustrated in section 2.2, a comparison of most biometric approaches is presented in Table 2-1. Table 2-1 also presents an analysis of how well the behavioural methods perform based on all seven requirements. The 'H', 'M', and 'L' in the table represent High, Medium and Low respectively. The permanence of characteristic in linguistic profiling is poor because user behaviour is changeable over time; (i.e. behavioural aspects of the individual, for example: signature, voice, gait and linguistics); therefore, such characteristics need to be carefully considered.

| Biometric behavioural approaches | Universality | Uniqueness | Permanence | Collectability | Performance | Acceptability | Circumvention |
|---|---|---|---|---|---|---|---|
| **Signature Recognition** | L | L | L | H | L | H | H |
| **Voice Verification** | M | L | L | M | L | H | H |
| **Gait Recognition** | M | L | L | H | L | H | M |
| **Keystroke Dynamics** | L | L | L | M | L | M | M |
| **Linguistic profiling**[2] | L | L | L | M | L | M | M |

Source: Jain *et al.*, 2004

Table 2-1: A brief comparison of behavioural biometric approaches

## 2.6 Linguistic Profiling

Linguistic profiling is a behavioural biometric technique that tries to verify, identify and discriminate users based on their writing style. Many types of linguistic features can be profiled, such as lexical patterns, syntax, structure, content specific, character content or item distribution through a text (Abbasi *et al.*, 2005; Stamatatos *et al.*, 2001; Zheng *et al.*, 2006). Extensive research has been undertaken on both techniques and a number of studies have used these techniques. In order to better understand the nature of the field more widely, both techniques of author verification and identification are involved. The next section provides the background to both techniques.

### 2.6.1 Author Verification and Identification

Forensic science has faced many challenges in many different kinds of crimes, from physical crimes to computer mediated criminal activities. Recognition of suspects has become crucial for law enforcement due, in particular, to the anonymity that the internet and associated services provide. Authorship analysis

[2] Inserted by the author

can be viewed from two different perspectives. Firstly, author identification by finding the most likely author of a target document or post in question, given the samples of the writing of a number of authors. The primary goal is to determine the possibility that an author wrote the document or post in question; the author would be one of those whose samples were provided (Zheng, 2006; Nirkhi, 2015). Some researchers maintain that this kind of authorship identification may not be suitable realistically, and could be invalid. For example, if the number of suspected authors is large, conceivably the suspected authors are likely to number into the thousands. Secondly, there is no guarantee that the true suspect of an anonymous text is among the known suspects. Finally, the amount of samples collected for each suspected maybe be limited, and the anonymous document itself may be short and limited (Koppel et al., 2013).

Authorship verification entails checking if a target document was written or not by a specific person by investigating other pieces of writings from that person; it gives a binary answer of “Yes” or “No” to the question (Brocardo et al., 2014 a; Li et al., 2014). This is appropriate for investigation because the suspect's message on multi-platforms would be available in the datasets and, therefore, the user's stylometric features for a known platform can be compared to those from an alternative unknown platform.

Whilst the roots of author attribution can be found within a paper-based human-oriented approach, more recently, it is being used within an IT context, where the IT system performs an automated verification or identification of the user, making this a firmly biometric technique (Aljumily, 2017). As previously highlighted, the approach, when automated, is analogous to linguistic profiling.

Recently, with the rapid growth of electronic messaging platforms, the nature and need for author verification has expanded to electronic documents. Alongside the introduction of computing came the widespread use of the term stylometry. Stylometry is the study of all features of a document. Stylometry in the field of author verification is an old research area that goes back to pre-computer times, it is the study of linguistic style, including length, word choice, word count, syntactic structure and other related attributes (Abbasi & Chen, 2005).The research area was developed by some scientists to assist text analyses. Its establishment was inspired by an American physicist, Mendenhall, in the 1880's (Klaussner et al., 2015). Mendenhall suggested that authorial fingerprints can be determined by counting the number of letters used in words. The idea of counting text features was later expanded in the 1900's to include the length of sentences (Nirkhi, 2012). Currently, there are various stylometric features used, and using more features in combination with each other can allow for increasing the discriminatory potential (Olsson, 2004). Determining word frequencies is assumed to be an effective stylometry for author identification. In addition, a combination of word frequency, character-based and function word attributes is an effective approach to identifying the author of anonymous texts, such as the author of an Email message (Klaussner et al., 2015).

Stylometry can be used in combination with other evidence within a wider criminal investigation to decrease the possible number of suspects (Koppel & Schler, 2004). The technique is rarely used as primary evidence – due to the level of performance – but can be useful in recognising and narrowing down the pool of suspects that require further investigation. Stylometry is a widespread approach for authorship studies, as it can help in analysing a given text and referring it or

attributing it to the original author. Table 2-2 shows the five common stylometric features used in author verification (Abbasi & Chen, 2005).

Table 2-2: The five common stylometric features used in author verification

| | Category | Description |
|---|---|---|
| **Stylometric Features** | Lexical | Lexical features are based on the idea that text can be broken into tokens. Each token is used to represent a character or word based such as *sentence/line length* (Yule, 1938; Argamon et al., 2003), *vocabulary richness* (Yule, 1944), and *word-length distributions* (De Vel et al., 2001; Zheng et al., 2006). |
| | Syntactic | Syntactic features is about the use of *function words* ("and", "but", "on", etc.) or punctuation. Syntactic features are more effective in identifying author than lexical features (Zheng et al., 2006). |
| | Structural | Structural features is about how the text is structured and organised. For example, the use of fonts, sizes, colours and several files extensions (Abbasi& Chen, 2005). |
| | Content-specific features | Content specific features is about the choice of words within a particular domain? For instance, content-specific features as in computing field "Laptop" and "MacBook". |
| | Idiosyncratic or character feature | Idiosyncratic or Character Features is about misspellings, grammatical mistakes, or thoughtful selection of words by the author for example, centre (American style) or centre (UK style). |

## 2.7 Conclusion

Physiological biometric recognition is built on the basis of differences in the human body, while behavioural biometrics are based on the user's behaviour. Physiological biometric methods can provide high protection for a system, since they have strong unique biometric features that are difficult to change or fake and tend to remain unchanged over the time. In addition, less time is needed to capture the initial and reference template samples. On the other hand, behavioural biometric methods may be less accurate and unique, and they

require a longer time to extract the features of users, although they can be used as identification techniques to determine the authenticity of the user.

Biometric systems for the identification of authors' stylometric features have been utilised in many security systems to identify and verify the authenticity of a user or a suspect, and there are a number of applications for this within forensics and law enforcement (Forstall et al., 2016; Rocha et al., 2016).

Author verification, or linguistic profiling verification, is a form of biometric system that is used to verify users based on their writing style. Writing style is unique and behavioural, and expresses the person's way of writing; it plays a large role in the process of demonstrating the user and specifying the language of the person's profile. Many types of linguistic features can be profiled, and it is built on the assumption that people have a characteristic pattern of language usage, as a sort of "authorial fingerprint". Exploring author verification and deriving it from the length of the text; the features of the selected vocabulary; the sentences used, and the structure of the messages created on different types of messaging platforms, plays an important role in providing the reference template used in the process of user verification. All in all, many methods can be used to distinguish authors linguistically, one of which can involves using the features of the user from a single platform and then finding similarities on other platforms and applying the necessary comparisons in order to identify and explore common features on different platforms and to facilitate the process of identifying the user. Expanding on the ideas in this chapter, the next chapter contains a review of the literature on author verification.

# 3. Chapter Three: Literature Review of Author Verification

The increasing popularity and diversity of messaging systems today, such as WhatsApp, Facebook, Twitter, Email, WeChat, and LinkedIn, has enabled people to communicate with each other in a convenient way via their mobile devices or Internet connected computers (Marques, 2016).

Although a large volume of literature is available on author verification for long documents, little literature exists on using author verification for short texts and on different messaging systems; although there is significant content on author identification, which is included in this study in order to better understand the nature of the field more widely.

Most of the techniques used in this area have focused only on verifying users' stylometry in individual messaging systems (Layton et al., 2010; Fridman et al., 2013; Allison et al., 2008; Abbasi et al., 2005; Stamatatos et al., 2007; Koppel et al., 2013; Ragel et al., 2013). Some researchers in this field have focused only on using the relationship between the same user's stylometry linked to different messaging systems through a technique known as "linkability" (Almishari et al., 2014 a); for example, linking the user's stylometry based on a user profile. Meanwhile, other research has focused on using techniques such as statistical analysis Mendenhall,1887; Farringdon,1996 (Fourkioti et al., 2019; Neal et al.,2017;Nagaprasad et al., 2015),mining users, clustering and classification (Iqbal et al., 2010 a).

Moreover, a critical evaluation of the literature on author verification in various messaging systems is essential to establish a higher degree of understanding of the domain. Consequently, this study will evaluate the work that has been carried out so far to better understand the most frequently used techniques, processes

and methods; analyse the results that have been developed, and discover the main challenges and barriers that have arisen with these systems. Finally, this literature review seeks to understand the potential feasibility of author verification of electronic messages within multiple messaging systems.

## 3.1 Literature Review of Author Verification

According to Nirkhi (2012), stylometry has a long history, with the first efforts dating back to the 18th century by English logician Augustus de Morgan, who proposed the possibility of exploring authorship by looking at whether one message consists of more long words than others. His suggestion was studied by Mendenhall in (1887), whose results on authorship attribution were successfully published and extended by others, including Bacon, Marlowe and Shakespeare (Nirkhi, 2012). The most detailed study on this topic was undertaken by Mosteller and Wallace (1964) who studied the ambiguity of the authorship of the Federalist Papers. Bayesian statistical analysis of the frequencies of a small set of common words was the first computational method used to guess the author of a text ("and", "to") in order to do discriminate between candidate authors. All 12 papers were attributed to Madison. Their results were later recognised and confirmed by historical scholars and thus became the first example of the approach being used.

In previous studies, features like sentence length (Yule, 1938) and vocabulary richness (Yule, 1944) were given due recognition (Barrón-Cedeno et al., 2017). In addition, subsequent research by Burrows (1987) came up with a set of more than 50 high-frequency words that were tested on the Federalist Papers (Neal et al., 2017). Holmes (1998) studied the use of "shorter" words (i.e., two- or three-letter words) and "vowel words" (i.e., words beginning with a vowel). Although

they started with the study of short texts, they did not consider or even find a specific mechanism for measuring the performance of identification.

Most of the analytical tools used for authorship analysis in earlier studies were statistical univariate methods. For example, the use of histograms of word-length distribution by Mendenhall (1887); characterising the stationary distribution of words or letters using the classifier Naïve Bayes (NB) by Mosteller and Wallace (1964); and the CUSUM (or cumulative sum control chart) statistics tool by Farringdon (1996) (the CUSUM is a sequential analysis technique that is used for monitoring change detection) (Nirkhi, 2012). According to Zheng (2006), the CUSUM statistics tool is used to produce the cumulative sum of the deviation of the measured variable to compare amongst users, and it is also employed as a forensic tool to help experts to confirm authorship analysis. However, there have been a number of failures because it was found by Holmes (1998) that CUSUM analysis is unreliable for forensics, since its stability over multiple topics is not good.

With the beginning of the use of computers, more widespread use of machine learning techniques were introduced (Argamon et al., 2003). Based on most experimental findings and results, it has been concluded by most scholars that machine learning methods are more accurate than statistical approaches (Nirkhi et al., 2012). According to Nirkhi et al. (2012) the performance of authorship analysis can be largely influenced by stylometric feature selection, which is used for writing style, in order to discover and find the most effective discriminators.

According to Brocardo et al. (2014), authorship analysis can be seen from three different perspectives. Firstly, author identification, including finding the author of the document or post in question, having been given the samples of the writing of

a number of authors; the main goal is to conclude which author wrote the document or post in question, and the reliability would be based on the samples that were given (Zheng et al., 2006). Secondly, authorship verification involves checking if a target document was written or not by a specific person by investigating other pieces of writing from that person; this gives a binary answer of “Yes” or “No” to the question (Brocardo et al., 2014; Li et al., 2014). Thirdly, concluding the characteristics of an author and deciding the author demographic (e.g. age, gender, race, culture, education, etc.). Table 3-1 shows the main fields and their tasks in the classification for authorship analysis.

Table 3-1: Classification and taxonomy for authorship analysis (Brocardo et al., 2014)

| Category | Description |
|---|---|
| Authorship identification or attribution | Deciding the most likely author of an anonymous document by comparing it with known existing documents. |
| Authorship verification | Checking whether a target document was written or not by a specific author. |
| Authorship characterisation or profiling | Concluding the characteristics of an author and deciding the author’s demographic (age, gender, race, culture, education, background etc.). |

As reported by Kebede et al. (2015), all the above sub-fields of authorship analysis are powerful enough to distinguish a single author from multiple authors by examining the stylometric features.

### 3.1.1 Stylometric Features

The majority of previous studies to date have focused only on stylometric features in order to achieve the recognition of authors. In fact, knowing the best set of features to be used in author verification can be a difficult task. However, the majority of researchers have combined two or more types of stylometric features (Abbasi et al., 2008). For example, they have combined syntactic features with lexical (Tan et al., 2010). Abbasi and Chen (2008) emphasise that there is an

urgent need to use larger feature sets that consist of various groups of features, for example, punctuation with word-length distributions, and combining lexical with syntactic, and syntactic and other features. According to Abbasi and Chen (2005) and Zheng et al. (2006), the use of feature sets containing lexical, syntactic, structural and context-specific features are more effective and operational for online recognition. The stylometric features are widely used, and some previous studies have indicated that many courts of law permit it as evidence in some countries such as the UK, the United States and Australia (Altamimi et al., 2019).

Abbasi et al. (2008) claim that stylometric analysis techniques can be classified into two main groups: supervised and unsupervised methods. Supervised techniques refer to the methods involving author-class labels for categorisation. This can be used in investigation fields such as forensics in order to identify criminals, as investigators typically have a limited number of suspects and so can identify and capture verified samples to compare with the suspect's messages (Mariappan et al., 2016). Whereas unsupervised techniques can be used where there is a lack of any prior knowledge of author classes (Khanum et al., 2015). Common supervised techniques applied in authorship analysis are support vector machines (SVM) (Nirkhi, 2019); neural networks (Zheng et al., 2006); Decision Trees (Abbasi et al., 2005), and linear discriminate analysis (Baayen et al., 2002), while unsupervised stylometric categorisation techniques consist of principal component analysis (PCA) and cluster analysis (Holmes, 1992).

There is another commonly known way to identify users, which is referred to as 'writeprints' (Abbasi and Chen, 2008). This technique represents an author's writing style, which is frequently consistent across his or her writings. These features are gathered from previous works and contain lexical, syntactic, structural, context-specific, and idiosyncratic features (Overdorf et al., 2014).

Another development in the field of stylometry is a technique called Doppelganger Finder (Afroz et al., 2014). This method was specifically designed to link users with multiple accounts within the same forum and has a rather complex framework in terms of how it is implemented (Overdorf et al., 2014; Greenstadt et al., 2014). Most of the methods used in previous studies to identify the author of a text involve mixing different features of stylometry in the recognition of authors through their text messages, as well as most previous research emphasising the importance of combining the features of stylometry in order to achieve high accuracy. The following sections will review the methods used in previous studies into the use of stylometric features for both long and short text messages.

### 3.1.2 Stylometric Features in Long Text

Long text refers to a greater number of words size in a document, such as books, articles, novels, online blogs or electronic forums. In previous studies on long text, the minimum number of words for long text was found to be 50 words for an effective study (Corney et al. 2002), while, the maximum number of words was found to be more than hundreds of thousands, as shown in Table 3 2. There have been many studies that have sought to enhance the performance of author attribution based on long documents, with accuracy rates of between 70% to more than 90% (Monaco, 2014), as shown in Table 3-2. As previously highlighted, many early studies combined two or more types of stylometric features such as lexical features (for example, word or character occurrence) and syntactic features (such as function words). More recently, renewed focus has been given to identifying different features (Narayanan et al., 2012). For instance, Baayen et al. (2002) used 50 common function words and eight punctuation symbols and tested these using 72 articles written by eight authors with 908 words per article. Their method involved measuring the degree to which non-professional authors

with a similar background can be distinguished on authorial structure in texts, and their results show this to be true, with an accuracy rate of 88.1%.

With the same objective of using a combination of stylometric features, a study by Zheng et al, (2006) focussed on the identification of online text messages. They demonstrated a method which can be used for multiple-languages. They tested their method on both English and Chinese newsgroup messages. For the English language, they collected an average of 48 messages from each of the 20 authors, with an average message length of 169 words. The same authors also tested their experience on the Chinese language, with an average of 37 messages per author. Each of these messages had an average of 807 words; the average message length in the Chinese language is longer than the average message length in English due to Chinese being a typical Oriental language which has no word boundaries (Zheng, 2006). A combination of 270 features were examined, as follows: 87 lexical, 158 syntactic (with 150 function words), 14 structural, and 11 content-specific. The experiments involved the use of three natural language classifiers: SVM, C4.5 decision tree, and back propagation neural network. SVM gave the best results, with a 90-97% accuracy rate for the English data set and 72%-88% for the Chinese data set. In both languages, SVM outperformed both the decision tree and neural network, as the SVM classifier has the ability to handle large-scale classification in long texts. Structural features and content-specific features demonstrated better performance in terms of discriminating capabilities for authorship identification on online messages, since they show how the author builds the content of a message structurally. In addition, content-specific indicates the level of depth of the author's cultural or other domain; for instance, the word "software" is always used by computer students.

In the same context, Iqbal et al. (2010a) studied 292 stylometric features, involving lexical, syntactic, structural and topic specific, with 158 users and 200 Emails per user. Their study focused on confirming whether a given suspect is the true author of a doubtful textual document or not. They used two Email datasets - one taken from a large population and the other one taken from a potential suspect, and each Email was converted into a vector of stylometric features. Their method involved verification using the Bayesian Network classifier. They claim that authorship analysis outputs for Emails of less than 500 words would not be significant in terms of improving performance. A limitation is that the style variation of the same suspect when he writes may affect his representative model, and they achieved an accuracy rate of 80.6%. However, combining lexical features with syntactic features and structure seemed to give the highest accuracy for Emails, because Emails are often formal and contain information addressed to the receiver of the message. Thus, the features of the structure and the diversity of words in the text indicate that the lexical and the syntactic features are all active, because most Emails contain information or explanations such as function words ("in", "or", "at", etc.) or punctuation.

Monaco et al. (2013) examined 30 book authors using 228 lexical and syntactic stylometric features. Each author had 10 books, and each book contained around 10,000 words. Their stylometry system used the following stylometric features: 49 character-based, 13 word-based, and 166 syntax-based features. In addition, the features were selected to show reasonable variation over a population of authors. For example, some authors use a large range of vocabulary and others use a small one. The features have been normalised, for instance the amount of different vocabulary, and the number of words. The 300 text samples were cut into files of eleven different sizes (250, 500, 750, 1000, 1500, 2000, 2500, 3000, 4000, 5000 and 10000 words) to obtain system performance as a function of text

length. They used the K-Nearest Neighbour (KNN) as the classifier, and achieved a 91.5% rate of accuracy in authorship authentication. Therefore, this shows that the selection of lexical features and syntactic features, and their integration with each other, gives highly accurate verification for long texts, especially for documents such as books. However, the differences in the number of users and the number of books may have had an impact on increasing the accuracy.

In the same context, using books with a different classifier, Koppel et al. (2004) achieved an accuracy of 95.7% when they used 250 of the most commonly occurring words that are often named n-gram to authenticate an author from among 10 authors using 21 books with texts of varying length for each author, and using SVM as the classification method. It is important to clarify what N-gram is: N-gram can be used for text mining, and "The approach is often called 'bag of words' because it simply counts word occurrences and mostly ignores word order" (Schonlau et al. 2017). Thus, the concept of an n-gram involves calculating the tokens of characters or words in the documents being considered in a continuous sequence, for example, word and character n-gram frequencies.

This indicates that the use of n-gram for determining the most used words also has an effect on verifying authors of books and long documents, since it displays the number of words or characters used in the text. However, both studies (Monaco et al., 2013 and Koppel et al., 2004) used lexical features to verify the writing style of the authors of books and long documents, and the study conducted by Koppel et al., (2004) showed a significant increase in accuracy because they used a SVM classifier rather than a KNN classifier, which indicates that SVM outperforms the KNN classifier, especially for long text.

Stamatatos (2007) used common n-gram features for author identification for when limited samples exist for testing; they investigated the class imbalance

problem and conducted an experiment on the compensation of imbalanced data sets. Data from 50 authors was collected from the Reuters Corpus Volume I (RCV1). Each author created 100 messages which ranged from 288KB to 812KB (a 1KB plain text file would hold roughly 200 words), and an altered Common N-Gram (CNG) method was used. They achieved an accuracy rate of approximately 70% by using SVM as a classifier. Other studies on author attribution for imbalanced data have proposed using many short text samples for the minority classes, and less and longer text samples for the majority classes (Vorobeva, 2016). Therefore, it is necessary to take into account the size of the samples used for testing to evaluate the accuracy when performing the identification process, and most studies in this area have proven that there is no mechanism solid enough to provide the appropriate size of samples for the testing to be applied in biometrics during the identification process.

Koppel et al. (2003) attempted to identify authors by using a combination of function words, bi-grams, and 99 idiosyncratic features, such as sentence fragments, wrong vowel and mismatched tense. Their aim was to use syntactic information based on syntactic error and evaluate the effectiveness of such features, both in and of themselves, and in combination with other types of features. They performed the study on 480 Emails from 11 authors during a period of nearly a year, and each Email included around 200 words. Three classes of features were used: lexical (i.e. function words: “and”, “the”, “that”), Part-of-Speech (POS) Tags (i.e. verb, noun) and idiosyncratic (i.e. syntactic, formatting and spelling usage). The appearance of functional words can be used as a marker for writing style and could be an indicator of authorship. The POS tagger was employed to the corpus to label each word with one of 59 POS tags, and after that, the frequencies of all POS bi-grams that appeared at least three times in the

corpus were used as the POS feature set. The study showed that the use of idiosyncratic features significantly improved the accuracy rate from 61.7% to 71.8%, using a decision tree as the classification method. In addition, it can be compared to the study by Iqbal et al. (2010a), as mentioned above, since both used Email as the platform. Although the number of users in the study by Iqbal et al., (2010a) are more than the number of users studied by Koppel et al., (2003), the use of lexical and syntactic features in the verification/identification of the Email author is more effective than the use of idiosyncratic features, because it is possible that the idiosyncratic features changed during the writing of the Email such as (English - British) to (English - American). For example, the idiosyncratic features of the word “center” instead of “centre”; the word “center” is frequently used in the United States of America, while the word “centre” was originally used in the United Kingdom.

Several researchers have examined the effectiveness of stylometry for authorship authentication and identification with text in the range of 75 to a few hundred words. For instance, Orebaugh (2006) developed an instant message intrusion detection system framework in order to test the instant message conversation logs of four users, based on 69 stylometric features, focusing mainly on examining character frequency as a stylometric feature, with some additional stylometric features, including: sentence structure, predefined specific characters, emoticons, and abbreviations analysis. The study was an attempt to analyse 2500 characters, which is 500 words, assuming that (1 word = 5 characters). The naive Bayes classifier was used, and it achieved an accuracy rate of approximately 68%. The results show that uppercase characters, special characters and numbers are distinguishable, and can be used as a form of intrusion detection system. According to Ali (2011), identifying and showing these features are the main

challenge for authorship identification, since they can contain emoticons, special characters and uppercase or lowercase letters.

In the same context of using limited words for gender identification which is a branch of the authorship problem, Corney et al. (2002) investigated four users; each user had 253 Emails and messages ranging from 50 to 200 words per Email. They used function words, structural, stylistic, gender attribute features and SVM for the classification, and they achieved an accuracy rate of approximately 70.2%. Their approach distinguishes between male and female authors, and the main finding is that function words provide the most important aspect for discriminating gender (Corney et al., 2002). Cheng (2009) observed that gender identification problems can be treated as a binary classification problem, such as class (1) for male and class (0) for female.

Generally, for most previous studies on long texts, especially books, online articles, electronic forums and Email, the most common stylometric features used for authorship studies have been lexical (such as word or character frequency) and syntactic (such as function words or punctuation). One of the most significant findings from previous studies on long documents is that the authorship attribution problem has been significantly influenced by using a combination of two or more types of stylometric features. Among the various combination features used, a combination of lexical with syntactic may be the best approach to identify authors in long documents, and it may be more applicable. This is because of the variety of words used in long messages, and to explain the message according to the words used (“but”, “although”, “at”, etc.).

Another significant finding from previous studies on long documents is that authorship attribution has been mainly influenced by the machine-learning paradigm. Among different classification techniques, the SVM and Bayesian

classifiers were regularly used. The SVM classifier seems better than the Bayesian classifier and decision trees, and Bayesian seems to outperform decision trees. In general, the performance of an accuracy of different types of long documents achieved an accuracy rate of 70% to more than 90% for 50-200 words. Table 3-2 categorises the previous studies according to the techniques, type of features and classification used, and the accuracy and analysis is presented as well.

As shown in Table 3-2, for long texts studies, the performance can be considered good as long as it deals with long texts such as books and blogs, because long texts make it easier for the classifier to achieve a good result, since the volume of information contains the full linguistic characteristics; unlike small amounts of text that deal with a limited number of features, which is difficult for most classifiers. Although the literature review has shown that the performance for long documents seems to be good, it is necessary to look at the volume of information being used in order to further determine this to ensure a good level of recognition. Whilst there is clearly something useful from language being identifiable, the literature review shows that the volume of words being used to achieve good performance far exceeds what would be expected from modern electronic systems

Table 3-2: The summary of literature review of stylometry with long text

| Author | No. of Authors | Samples each suspect | Sample Size | Feature Types | No. of Features | Classification type | Accuracy | Goals of study |
|---|---|---|---|---|---|---|---|---|
| Zheng et al. 2006 | 20 | 48 online forum postings | 169 | Lexical, structural, Syntactic, and content specific | 270 | SVM decision tree, and NN | 97.69%for SVM, 96.66% for NN and 93.36%for C4.5 | Identification |
| Tan, et al. 2010 | 2 | 167 blog posts | 170-357.5words | 13 Syntactic and 4 lexical | 21 | Naïve Bayes | 81.98% | Identification |
| Steyvers, et al. 2004 | 85 | Average 1882 abstracts | differs | Author- topics and topic-word models | 300 | SVM | 72%, | Topic discovery |
| Stamatatos 2007 | 50 | 100 | 288KB 812KB | Common n- gram | Not specified | SVM | 70% | Identification |
| Pavele, et al. 2009 | 20 | 30 short articles | Not specified | Conjunctions and adverbs | 177 | Prediction by partial matching (PPM), and SVM | 83-86% for PPM 82.9-84% for SVM | Identification |
| Monaco, et al. 2013 | 30 | 10 books | 10000 words | Lexical and syntactic | 228 | K-NN | 91.5%, EER 8.5 | Authentication |
| Koppel, et al. 2004 | 10 | 21 books | About 500 words per chunk | Common words or partial word (n-gram) | 250 | SVM | 95.7% | Authentication |
| Iqbal, et al. 2010a | 158 | 200 Emails (Enron corpus) | <500 words | Lexical, syntactic, and structural | 292 | Bayesian network | 80.6%, EER 19.4 | Authentication |

| Author | No. of Authors | Samples each suspect | Sample Size | Feature Types | No. of Features | Classification type | Accuracy | Goals of study |
|---|---|---|---|---|---|---|---|---|
| Corney, et al. 2002 | 4 | 253 Emails | 50-200 words | structures, stylistic function words and gender- attributes | 222 | SVM | 70.2% | Gender discovery |
| Baayen, et al. 2002 | 8 | 9 fictions | Average 908 words | 50 Function words, 8 punctuation | 58 | Entropy-weighted linear | 88.1% | Identification |
| Orebaugh 2006 | 4 | 35 Segmens of instant messages | 2500 characters=500 words. | Sentence structure, emoticon, abbreviation. etc. | 69 | Naïve Bayes | 99.29% | Identification |
| Howedi,et al. 2014 | 10 | Average 3 text message | 290-800 word | Lexical, structural, Syntactic, and content specific character N-gram | Not given | Naïve Bayes and SVM | 96% | Identification |

In most previous studies on long documents, the minimum number of words was found to be 50 words, as shown in the study by Corney et al. (2002). In addition, most previous studies have treated stylometric features in a certain way by trying to combine linguistic features with each other; therefore, linguistic characteristics differ from one study to another, even within the same platform and area; for example, the studies by Monaco et al. (2013) and Koppel et al. (2004) both used the same platform (books). This indicates that most of the features examined in previous studies have been treated in a certain way through the combination of lexical, syntactic, structure and so on, and this indicates that it is not clear which linguistic characteristics can achieve good performance for individual platforms, or whether it is possible for one feature (e.g. lexical) to be more reliable than the other on a single platform or several platforms.

As it can been seen in Table 3-2, verification studies and identification studies have used stylometry. Most previous studies have focused on identification techniques, which involves finding the author of the document or post in question. Given the samples of writing of a number of authors, the goal is to determine which author wrote the document or post in question. The author would be one of those from whom samples were provided. While the authentication technique, also referred to as authorship verification, involves using a document or a post to determine if it was written by a specific user. However, none of these studies on long documents have identified or explored common features across different platforms and different domains, rather, they were conducted using only a single corpus.

### 3.1.3 Stylometric Features on Short Text in Messaging Systems

This section focusses on studies that have sought to specifically use short messages. Short text messages are defined in most previous studies as

containing 75 words or less (see the comparison features in Table 3-3). Instant messages, text messages, and social network messages are typically shorter messages, unlike other online messages or posts such as blog posts or online articles.

As can clearly be seen above in the stylometry of long texts, the traditional stylometric features, particularly lexical and syntactic features or a combination of both of them, are more applicable for single platforms. However, with microblogs or social network messaging systems, users can simply post a message as a quick update of their status or the activity they are involved in. Twitter is one of the social networks that places a restriction on the amount of text, which restricts its users to a maximum of 140 characters. Therefore, certain stylometric features that have been used for identification techniques, such as structural features, may not be applicable or effective because users do not have much control over the content of the post, and could potentially modify their behaviour in order to conform to the restrictions placed on them by the messaging platform.

Table 3-3: Summary of the literature review of messaging systems in identification studies

| Author +Year | No. of Suspects | Samples for each suspects | Sample Size | Feature Type | No. of Features | Classification | Accuracy |
|---|---|---|---|---|---|---|---|
| Layton,et al. 2010 | 50 | 120 tweets | 140 Chars max | Character n-grams | Not specified | SCAP algorithms. | 70% |
| Green, et al. 2013 | 12 | 120-900 tweets | 140 chars max | Bag-Of-Words and style markers | Hundreds features of bag- of-words and 86 style | SVM | 12 users were tested gain 40.5%, |
| Allison,et al. 2008 | 9 | 174-706 Emails (Enron corpus) | 75 words | Word frequency, 2-grams, 3-grams and stem words | Not specified | Multimodal Hierarchical SVM | 78.46% 87.05% 86.74% |
| Zheng, et al. 2006 | 20 | 30-92 Emails | 84-346 words | Lexical Syntactic, structural | 270 | SVM and C4.5 | 97.69% and 93.36% |

Layton et al. (2010) tested 50 Twitter users with each user, having 120 Tweets. They used a 3- gram approach and the Source Code Authorship Profile (SCAP), and the study obtained an accuracy of 70%. The users of Twitter can use a "#" followed by a tag name to link messages with specific topics. Also, users of Twitter often use "@" followed by a specific user's name to direct the message to a destination. All these structural contents count toward the 140 character limitation; however, all these structural contents were removed by the researchers before applying the SCAP algorithm to allow SCAP to focus on the actual content that the user wrote. Furthermore, the SCAP method extends the work by Keselj et al. (2003) on classification (Layton et al., 2010). In Keselj's study, an author profile is described as "a set of length L of the most frequent n-grams with their normalized frequencies." and an n-gram is the number of characters in a continuous sequence. The profile of an author can be indicated as {(x1; f1), (x2; f2)… ($x_L$; $f_L$)} for i=1…L, where $x_i$ indicates to an n-gram and $f_i$ indicates to the normalized frequency of $x_i$. For SCAP, the frequency $f_i$ of the n-gram $x_i$ was not normalised. A profile of an author is defined as the L numbers of n-grams which have the highest frequency, which can be: {x1, x2….xL}.

Throughout the classification process, and when an unknown profile was accessible for authorship identification, the author who shared the most n-grams with the unknown profile will be specified as the author of the unknown profile. However, the drawback to their method is that an increase in messages would not have any further positive effect on the accuracy. In addition, their method is questionable in relation to the authorship identification task of not so common messages, as the accuracy rate dropped by 27% when data about the discusser's user information was taken out (Rappoport et al., 2013). One of the most significant techniques used to identify the author on Twitter is to use n-gram to

create a reference template that contains a continuous sequence of *n* items of a particular sequence of text, which has the power to collect and distinguish the characters of Twitter as long as the limit of characters on Twitter is 140.

Similarly, Green et al. (2013) studied authorship identification on Twitter, and collected data from only 12 users, with 120- 900 tweets per user. The feature set used comprised of Style Markers and Bag-of Words (BOW). The number of style markers used was 86, and these contained punctuation, long words, part of speech, hyperlinks, and other similar attributes. The BOWs contained all the words that came from the raw data, which was used as a measure when the words appeared more than five times in the whole dataset. SVM were used as a classifier. They found out that Style Markers performed better than BOWs for short text, with an accuracy ranging from 60% to 76.75% for BOWs, and 75.1% to 92.3% for Style Markers. The drawback is that when the researchers examined the effect of the number of authors, they found that the accuracy decreased from 92.3% when two more authors were added to become 40.5% with 12 authors. The reason for the low accuracy rate is that the greater the number of users added, the lower the accuracy rate. The increase in the number of users has a significant impact on the parameters process, especially with the Style Markers. Because each new user seems to have new patterns of writing style, that affects the stability of the overall measurement specified in the balancing parameters. Features of style markers may be suitable for identifying authors in small samples of datasets because it may be weak when the number of authors is increased (MacLeod et al., 2012).

In comparison to the previous study, it can be inferred that the accuracy of the study by Layton et al. (2010) dropped to approximately 27%, while the study by Green et al., decreased in accuracy by approximately 51%, taking into account

that the number of authors in Layton et al is four times larger in comparison to the number of authors in Green et al., which indicates that n-gram can play an important role and is more effective for Twitter. Because the n-gram has the ability to handle characters, as well as the ability to distinguish them within text, this leads to user identification.

Allison et al. (2008) focused on author identification for Email. They investigated nine users of short Emails, with approximately 75 words each and with a range of 174 to 706 Emails per user; Enron Email corpus was used, along with 2-grams, 3-grams and word frequency measures, they derive an explicit estimate for the probability which a new document be appropriate to each of the likely classes, No. of features is not specified. SVM was used as the classification engine, which produced around 86.74% accuracy.

In comparison with the previous studies mentioned that used the Email platform for short text. Allison et al. (2008) and Corney et al. (2002) both utilised SVM as the classifier. Allison achieved an identification rate of 86.74% while Corney et al. (2002) achieved 70.2%. It should be borne in mind that the message length of Allison is 75 words and Corney's message length is between 50-200 words, and the differences are only in the number of users and the number of chosen features in stylometry. This indicates that there is a weakness in determining the best features to be used in stylometry, the best size Email message, and the appropriate number of users on the Email platform in the identification process. This led to discovering the optimisation of features mentioned above, because there is no solid basis for reliance on during the investigation or exploration.

Koppel et al. (2007) and Sanderson et al. (2006) investigated 500 words from book and newspaper journalists, respectively. They used an approach called "Author Unmasking". The idea of author unmasking is that the differences

between two texts from the same author will be reflected in a relatively small number of features. These features can be extracted through the use of an author unmasking curve. In Koppel et al's experimental, they selected the 250 most frequent words from a collection of 21 English books published in the Nineteenth Century. These books were written by ten different authors. Each book was divided into chunks of equal sections of at least 500 words without breaking up paragraphs. An SVM was used for cross-validation. They obtained an overall accuracy of 95.7%. In Sanderson et al's experiment, they used 50 newspaper journalists, with a minimum of 10,000 words per journalist, and they divided the training set into 500 characters per chunk used. A SVM classifier was used, and the accuracy achieved was over 90%. They conclude that measuring of the "depth of difference" between two example sets is a different type to other measures, such as margin width, which could be based on a single highly differentiating feature, but it is not appropriate for this measure to be applied to other applications. This because this method is more appropriate when the unknown texts are long enough in length to allow those texts to be segmented into multiple parts to train the classifier (Stamatatos, 2009).

Siham and Halim (2012) mentioned that the idea that the longer the text, the better the identification accuracy will be. The possibility of short document analysis is difficult to carry out since this type of document is usually characterised by poor structure and informal language, which is seen much less in literary texts (Brocardo et al., 2014 a). For example, the content of the short message that exists in social networks such as Twitter and Facebook is often written without the author's full consciousness; this is due to the fact that the writing practice provided by the platforms does not allow more content and may be restricted by

certain linguistic factors; thus, the text message may include unclear language and have an unorganised format.

In general, the researchers have not agreed on a clear vision and a general framework to be used when the message is poorly structured with informal language to determine its validity and its suitability when it is being used in the investigation process. For example, the length of the required message, weaknesses and strengths of stylometric features, as well as acceptable format.

Having said that, the number of research studies that have been conducted on short texts is smaller in comparison to the research into long texts. The majority of the studies on short text are associated with digital copies, for instance, Emails or social network posts, ranging from 140-characters for Twitter to Emails of 75 words. Lexical and syntactic features were commonly used in short texts, and SVM is a common classification method. Most of the previous studies have shown that the effect of lexical and syntactic is an important factor in short text messages for the following reasons: Firstly, short text messages are usually a summary and specific to be understood by the other party, with the inclusion of only a few words, so the author takes into account the impact of the words. Secondly, Short text messages contain rules for the sentence to be adopted by adding words such as “but”, "and","on","therefore" and so on, as well as punctuation. This plays an important role in the process of syntax in the text message because most texts contain a comprehensive and accurate explanation. Thus, for example, a Twitter message containing 140 or 280 characters, the purpose of which is to deliver text message to the other party in a conceptual and concise manner. Technically, the most common feature addressed to deal with short messages is n-gram because it plays a large role in determining the characters of letters-words in sentences.

As discussed above, most previous studies have focused on the n-gram and have achieved a high level of accuracy.

With respect to language, there is a limited volume of research, with only a few studies on the Greek, Arabic and Chinese languages undertaken (Zheng et al., 2006; Stamatatos et al., 2001). The majority of authorship identification research has dealt with English language attributes and identification methods. For instance, word-based lexical features (as in the number of words in a sentence) are relevant in the context of English writing, but this does not apply to some other languages such as Arabic or Chinese, since the Chinese language has no explicit word boundaries (Zeng., 2006). Stamatatos et al. (2001) studied the Greek language and report that Greek is closer to English since they have similar linguistic characteristics such as word boundaries. Abbasi & Chen (2005) also point out that there are a lot more words used in Chinese than English. For the Arabic language, there are 28 letters while in English there are 26. The suggests that vocabulary richness in terms of resulting features may vary between languages, and a solution in one-language may not map to other languages.

Unfortunately, most the previous studies have focused on identification techniques, and a small-scale classification problem with two or three authors, often using long text samples. Only a small amount of research has explored single authorship verification, and no any studies have been found on across electronic platforms verification. Despite attempts to search for research that is related to cross platform electronic electronic messaging systems for authentication problems, no such authentication research across modern platforms was found. Even for the authorship identification problem that has been tackled by many researchers, there is only a very small amount of research that

has been performed on single platform authentication. Therefore, the next section discusses author verification on single messaging systems in more depth.

### 3.1.4 Author Verification on Messaging Systems

In line with previous studies, and in terms of verification on most single social messaging platforms, Table 3-4 shows the most recent studies conducted for different messaging platforms. Broadly speaking, little research has been found on stylometry across many of the modern platforms. The seminal work in this field was conducted by Brocardo et al., (2017). Their research study achieved an EER of 16.73% for 10 users and 100 samples per user. Lexical, syntactic, and application-specific features were utilised in the features set. Their technique relied on an n-gram technique to measure the degree of similarity between a block of characters and the profile of a user. On the Text message platform, the seminal work was conducted by Saevanee and Clarke (2011), and their research study achieved an EER of 24%. These findings are based on 30 participants, with a minimum of 15 samples per user; maximum samples was not mentioned and a Radial Basis Function (RBF) neural network classifier was used. The EER was 24%, and several users experienced an EER of 0%.

The most prominent previous study of the Facebook platform is by Li et al., (2014). They used posts from the Facebook platform to determine whether a user can be verified from among 30 users. Furthermore, they used SVM Light as the classifier, with 233 features; a total of 9259 posts were applied and 12 tests were conducted. For 10 users with 233 features, they achieved an accuracy rate of 81.6%. When the author number was increased to 20 and 30, the success rate dropped slightly to 79.8% and 79.6% respectively, with an EER of approximately of 20%.

For Email, the most prominent previous study was by Iqbal et al. (2010a), which yielded EERs ranging from 17.1% to 22.4%. The approach taken in their study was to cluster the anonymous Email using stylometric features and extracting the 'writeprint' to verify the author. They extracted 292 different stylometry features from 158 users and then analysed these features.The experiment is evaluated by using three clustering algorithms: Expectation Maximization, k-means and bisecting k-means, and achieved an EER of 17.1%. However, their technique was based on clustering and mining the writing styles from a collection of Emails written by multiple anonymous authors, and they attempted to group Emails written by the same author. The Enron dataset was utilised, which has been used extensively for authorship analysis research under a variety of different methodological methods, including text categorisation (Neal et al., 2017). Another prominent previous study on Emails verification (Brocardo et al., 2014 a) yielded an EER of 14.35% using an n-gram technique and the Enron corpus involving 87 authors. They used two steps: in the first step, the user profile was derived by extracting n-grams from sample documents. In the second step, a user specific threshold was computed and used later in the verification phase.

Table 3-4: Summary of literature review for messaging systems in verification studies

| Study | Document Types | Authors /#Texts | Method/ Approach | Feature type | Classifier | Performance % |
|---|---|---|---|---|---|---|
| (Brocardo et al, 2017) | Twitter | 10/100 sample per user | N-gram | Lexical syntactic structural | Gaussian-Bernoulli | 16.73% (EER) |
| (Saevanee et al, 2011) | Text message | 30/ Min15 sample per user | Calculate user word profiling & linguistic features | Lexical Syntactic Structure emotional keywords | Neural network (RBF Classification) | 24% (EER) |
| (Li et al, 2014) | Facebook | 30 | Compare classifiers SVM and C4.5 | Lexical syntactic structural short messages features | SVM | 79.6% (Acc.) (EER) ≈ 20% |
| (Iqbal et al, 2010 a) | Email | 158/ Enron email 200,399 e-mails | clustering | Lexical syntactic Structural & content-specific | EM, k-means, and bisecting k-means | 17.1% - 22.4% (EER) |
| (Brocardo et al, 2014 a) | Email | 87/Enron email 200,399 emails. | n-grams | Lexical syntactic structural & content specific | Ad hoc similarity. Distance(Percentage of shared n-grams) | 14.35% (EER) |

Unfortunately, the literature on author verification on messaging systems has focused only on single platforms, and with limited datasets. There is also a lack of analysis of the underling feature vectors that are appropriate for users within and across platforms. The next section demonstrates some of the studies that have attempted to connect users across social media sites.

### 3.1.5 Connecting Users Across Social Media Sites

This section will discuss the previous studies that aimed to connect users across social media sites with the aim of identifying and/or tracking the accounts of the same user across different social platforms. This has currently been attracting an increasing amount of attention and effort due to the significant research challenges and the vast practical value of the problem.

Moreover, the writing habits of online users can be used to create an author "writeprint" that can be utilised for their verification. This means that some unique

features such as structural layout behaviours, unusual language usage, and sub-stylistic features can all contribute towards helping create an appropriate feature collection or stylometric behaviour profile of users. Therefore, most of the above-mentioned studies have used this for a single and specific platform, while a small number of researchers have expanded it to include multiple platforms, either with a clustering technique or by tracking users.

A study by Novak et al. (2004) researched the issue of “anti-aliasing”, which attempts to identify unique users from among a set of online pseudonyms, based on their online content. They suggest a language model-based approach for authorship classification on online forums, for example, to connect several aliases of known individuals by using their public postings online, such as bulletins, weblogs and web pages. They attempted to develop algorithms to anti-alias those users because they believe users on bulletins, weblogs and web pages can adopt multiple aliases. Their contribution is to establish data mining to match users by using clustering. They used clustering to address two problems: Firstly, the features of the content authored by an alias and the new content, and deciding on the likelihood that the alias created the new content. Secondly, using computing likelihoods to decide the most suitable clustering of aliases into authors. The procedure they used for matching the aliases for 100 authors was to split them into 200 aliases, meaning that the writings of 100 authors were each divided into two, before using algorithms to match for similarity; that is, agglomerative clustering algorithms from machine learning, which begins with a set of entities, for example, documents in their own cluster, and repeatedly agglomerating the two closest clusters into one. Specific feature sets were used to represent the texts, and they used a cluster technique for clustering into 100 pairs. The features include word/vocabulary, misspellings, punctuation,

emoticons and function words. Their method may be functional based on the vocabulary used in the online postings, and they achieved more than 90% accuracy. Their limitations are concern over the development of the algorithm to be used when the number of authors is large, because there is no mechanism to develop it. Secondly, also related to their algorithm, it is not optimised to run at web scale. Their experimental results were satisfactory, although the similarity of aliases was directly linked to the topic written as the users' vocabulary was considered a discriminating feature. However, this research method might not be appropriate for people writing about heterogeneous topics. In addition, most of the writing style on bulletins, weblogs and web pages is presented as a reply to the same main topic or as comments to reply to the main topic. Also, there is no information available to determine the amount of posted content that can be used to identify users, for example the minimum length of post size for users, because most of the aliasing has to have a relation to the length of content for each user to be accepted into the identification process.

There is another technique that has been proposed and tested for social network analysis, which is called Hydra. The aim is to track users with multiple aliases on social media sites. It was introduced by Liu et al. (2014) and is a solution framework that allows large-scale social identity linkage through social media sites. In other words, it is a process for linking accounts of the same user across different social network platforms. In order to do so, the authors explain that there are three important problems that must be taken into consideration when linking user profiles or the same user across different platforms, which are: Firstly, completeness, as each platform is constrained by specific features and design, and specific orientation in its own style, and so the user profile will be segmented based on what is offered from the features of each platform. Secondly,

consistency, as the information provided by each user platform may be incorrect or incomplete, and so information provided by multiple platforms should help to develop consistent information about the user. Thirdly, continuity, as the user's identity remains over time, which makes it possible to integrate useful user information even when they become less popular. According to Liu et al. (2014), the Hydra in this approach involves combining heterogeneous behaviour modelling of user profiles through three stages: the first stage is to model heterogeneous behaviour through using long-term behaviour, including multiple resolution and matching time information; the second stage is to create structural consistency among the users to measure the level of steadiness of the platform's structure, and the third stage involves mapping functions through different objective optimisation processes.

In their procedure, they used five social networks that are popular in China, which are: SinaWeibo, TecentWeibo, Douban, Renren and Kaixin, and two worldwide social networks - Facebook and Twitter. Each user had accounts on every one of the five platforms. "Heterogeneous Behaviour Modelling" has been utilised to measure the similarities between two users during several phases. The first phase addressed user attributes that are either textual attributes, such as name, gender or age, or virtual attributes, such as face images that are used on user profiles. The second phase used the topics of interest of the user, for example, politics, religion, sport, and so on. The third phase considered the user's language style, such as individual words and emoticons. The last phrase examined information on the user's location and multimedia sharing, such as videos and images on the internet.

The main idea of Liu et al. (2014) was to create a linkage function via a multi-objective optimisation system. The system is based on the decision model on

pairwise similarity and users' social structure consistency information. The total score for the matching similarity process between users' behaviour captures the highly correlated actions between user accounts over a specific time; they then developed a linkage method by measuring the agreement of the social structure level behaviour. According to Goga et al. (2015), Jain et al. (2015) used attributes without analysing their features and their limits to match profiles in practice, therefore they used attributes with low availability, which can only match a portion of profiles across a small number of social networks and are likely to give many false matches in practice. This shows that the similarity factor in text messages may play a significant role, especially when the number of users is large. The authors used the "Heterogeneous Behavior Modeling" method in order to combine the heterogeneous behaviour modelling of user profiles and then to measure the similarity between each two users.

However, this might be a problem and there is a great possibility that the user could manipulate through his behavior as long as he has the ability to change the platform used, as whatever the reason for the change in platform, it is a change in behaviour; therefore it is necessary to know the least differences to determine the proportion of manipulation, and these are not available in this study. Secondly, monitoring a user's behaviour over a long-term period gives a high probability of recognition of the author of the message because the time factor plays a major role in this method. Therefore, it is necessary to determine the time taken in each process to connect the time with processes, and this is also not available in the study by Jain et al. (2015). Moreover, their study aims only to create a similarity between users without knowing the linguistic characteristics that influence this similarity. In addition, the study uses the name, age, sex and facial image from the user profile, but these do not explain the nature of linguistic characteristics,

despite the result of high performance because they introduced other factors such as age, gender, name and face image.

Another study conducted by Almishari et al. (2014 a) explored the linkability of tweets in order to link Twitter accounts. Their aim was to explore the stylometric similarities between multiple sets of tweets for the same author. Their procedure used two datasets, each containing over 8,000 Twitter accounts and an attempt was made to link Twitter accounts based on simple lexical features – unigrams and bigrams and combining hashtags with stylometric features to improve linkability. Naive Bayesian is the classifier engine used, and it was found that at least in the case of relatively active tweeters, and as far as the same author is concerned, linkability of tweets can be obtained without much difficulty, despite the large number of users, and an accuracy rate of nearly 100% was achieved. Although Almishari et al. (2014 a) achieved a high level of accuracy, and while two simple lexical unigrams and bigrams were used, the author did not address the size or the length of user messages. However, linking the amount of user messages is important to the investigation, and weakness occurs if the length of the user's message is not investigated, as it is not necessarily the case that every user on Twitter uses 140 or 280 characters in all their Twitter messages. Indeed, lexical features seem to have contributed to the increase in the accuracy rate, as the Twitter platform relies mainly on characters, vocabulary and expression in order to convey a message to the audience. According to Robinson et al. (2016), when President Trump writes a tweet, the words he uses the most are: bad, win, join, totally, people, and so on, which is mostly a lexical feature. However, their study only includes the single platform Twitter.

All in all, as can be noticed in the review of previous studies on connecting users, research into ‘identity recognition’ overlaps with many fields of science, such as

text mining and pattern recognition; in addition, each platform has its own special technique, with the different platforms also having a different technique to the other. The majority of it is related to user behaviour only on single platforms, and linking users within it, without mentioning which of the best features are most influential across modern platforms.

## 3.2 Discussion

This section is divided into addressing two core issues. The first discusses the mechanism of author identification and verification in long text messages as well as short text messages on single platforms. The second core issue discusses the mechanism of connecting author accounts within platforms.

This section will start by focusing on addressing the core issue of the identification of authors on single platforms for long texts, especially books, online articles, electronic forums and Email. A comprehensive discussion of a comparison between the previous studies, including number of users, types of features, number of words, performance and platforms will be presented. The highest number of users is in the study by Iqbal et al. (2010), which included about 158 users and is higher than all previous studies that used Email as the platform. However, their main limitation is that the style variation of the same suspect when writing may have affected this representative model. On the other hand, the smallest number of users (two) is in the study by Tan et al. (2010) who used 167 blog posts per user. Their limitation is that each of the two authors exhibited differences in the length of their entries on the database in terms of word count. The differences between the two aforementioned studies include the number of users and platforms, and while both of them used lexical and syntactic features, Iqbal et al. (2010) also used the structural feature, which seems to have increased the rate of accuracy and plays an important role in the Email platform, even if the

number of users is large. In addition, structural features may be highly useful because they reveal the manner of the user's writing style while using Email as a platform; however, in the study by Tan et al., they did not use this feature, although the number of users was less. The reason for the absence of this feature by Tan et al might be because writers of blog posts can write randomly without concentrating, and may not write official text messages such as Email, which may be why Tan et al. (2010) did not consider this feature.

Regarding, regarding Email, Corney et al. (2002) also used Email as a platform for identification. The number of users was four; the samples for each user were 253 Emails; the sample size was 50-200 words, and the feature types were structure, stylistic, function words and gender-attributes. The number of features was 22, the classifier used was SVM, and they achieved an accuracy of 70.2%. Their main limitation is that there is a need for a larger range of samples in order to increase the performance results. Although the study by Corney et al. (2002) and the study by Iqbal et al. (2010) both used Email as the platform, there are differences between them concerning number of users, feature types and classification. In fact, the function word feature in the study by Corney et al. (2002) played a major role in increasing the accuracy of the study, whereas Iqbal et al used the structure feature. However, the reason for the decrease in accuracy in the study by Corney et al. (2002) compared to the study by Iqbal et al (2010) is due to the use of samples of 50-200 words for each. Furthermore, the SVM classifier also contributed towards increasing the accuracy of the study by Corney et al. (2002). Despite the volume of the length of the data being small, as determined by the study (Corney et al., 2002), SVM has strongly contributed and can play an active role, especially with a small data size. This indicates that it has

outperformed the Bayesian classifier, which is why it may be the best classifier when a small volume of data is used, specifically with Email as the platform.

Studies using books were conducted by Koppel et al. (2004) and Monaco et al. (2013). The number of users in Koppel et al's study was 10 users; the samples for each suspect were 21 books; the sample size was 500 words per chunk; the type of feature used was common word (n-gram); the number of features was 250 features, and SVM was the classification type used. Their main limitation is that an unmasking approach might find more general application, although they achieved an accuracy of 95.7%. While in the study by Monaco et al. (2013), 10 books were used; the number of users was 30; the sample size was 10,000 words; the features used were lexical and syntactic; the number of features was 228 features, and the classifier was K-NN. Their limitation is that the database is relatively small, yet they achieved an accuracy of 91.5%. Although the sample size is bigger at 10,000 words compared to 500 words per chunk in the study by Koppel et al. (2004), the main difference is in the classification, as it can be noted that the SVM is superior and more effective than K-NN, as it makes it possible to identify the authors of the books. Moreover, the text has been chunked to form a smaller number of parts, even where the database is small; this is because it focuses on words and vocabulary, which may be the reason for the high level of accuracy in the study by Koppel et al., (2004).

In comparison, Zheng et al. (2006) studied online forum postings using three classifiers, which are SVM, NN and C4.5. Their limitation is that different parameter settings of authorship identification had an impact on performance. For example, the number of authors and the number of available sample documents in the training set. The average length of message per author was 169 words, and also less fewer words were used than in the study by Tan et al. (2010) mentioned

earlier in section 3.2.2. The feature types were lexical, structural, syntactic and content specific, and the number of features was 270 which more compare to the study byTan et al. (2010) . It can be noticed that the SVM Classifier outperformed Decision Tree, and that both studies (Tan et al., 2010 and Zheng et al., 2006) used lexical and syntactic features, although the study by Zheng et al. (2006) added two more features, which are content specific and structure. Moreover, all of the classifiers used by Zheng et al. (2006) outperformed on the classifier Naïve Bayes used by Tan et al. (2010), with a relatively large difference. Furthermore, SVM also outperformed NN and the C4.5 classifier. This gives an indication that SVM can play an effective role in identifying authors on online forums, since the SVM classifier has the ability to handle large-scale classification in long texts. From another perspective, the structural feature used by Zheng et al. (2006) also played an important role in the identification of the authors on online forums, since most of the forums have a special structure and users must follow specific procedures.

Some studies did not provide sufficient data and contain incomplete information for verifying the results, so these have not been included because most of their data are insufficient for comparison, although most of their indicators are fewer than in the studies described above.

In general, the most commonly used stylometric features for author identification are lexical (such as word or character frequency) and syntactic (such as function words or punctuation). One of the most significant findings in the previous studies involving long documents is that authorship attribution is significantly influenced by using a combination of stylometric features which combine two or more types, since this could help to improve the accuracy of identification. The longer the text is, the easier it is to compute stylometric features, which makes it more reliable

as more text is considered. From among the various combinations of the features examined, it seems that a combination of lexical with syntactic can achieve high accuracy in identifying authors of long documents and may be more applicable. As well as containing a range of vocabulary and characters. In addition, long texts contain function words such as “in”, “or”, “at”, and punctuation. By combining them with each other, it is possible to achieve higher accuracy compared to other features, as observed in most studies (Zheng et al., (2006); Monaco et al., (2013); Iqbal et al., (2010); Howedi et al., (2014)). The use of n-gram for determining the most used words is an effective way of verifying authors of books and long documents, since it shows the most frequent sound-oriented information in a text. For example, in function word features, the functional n-gram shows the elements of most of the lexicon in the text; examples of this can be found in the studies by Stamatatos et al. (2007); Koppel et al. (2004), and Howedi et al. (2014). A final significant finding in previous studies of long documents is that authorship attribution has been mainly influenced by machine-learning. Among different classification techniques, SVM and Bayesian are the most used classifiers. In addition, the SVM classifier seems better than both the Bayesian classifier and Decision Tree, and the Bayesian classifier seems to outperform Decision Tree. Overall, the performance for different types of long documents achieved an accuracy rate of 70% to more than 90% for 50–200 words. The study that achieved best for the minimum length and volume of words in long text is Corney et al. (2002) who used Email. Although Email has also been used on short text of 75 words in length in the study by Allison et al. (2008), but the difference is that Corney et al. (2002) used the length of the words as the variable, which is not specified by number, while the study by Allison et al. (2008) determined the results without any disparity, thereby distinguishing this study from the one by Corney et al. (2002).

The second part of this section will focus on addressing the first core issue concerning identification of authors, especially on single platforms for short texts. However, the amount of research is limited. The majority of studies on short texts have been applied to Emails or social network posts, ranging from 140-character texts for Twitter to Emails of 75 words. That is, except for the study by Zheng et al. (2006), as the texts range from 84 to 346 words, with samples for each suspect were between 30 and 92 Emails. This indicates that increasing the number of samples for each suspect for short text messages is also an important factor with regard to message size, because the greater the number of samples, the higher the recognition rate, and vice versa.

From another point of view, Layton et al. (2010) and Green et al. (2013) have both used Twitter as the test platformfor identifying users. The type of feature used in the study by Layton et al. (2010) was n-gram, although the number of features was not specified; the number of users was 50; the sample from each user was 120 tweets; the sample size was 140 characters maximum, and SCAP algorithms were used as the classifier. The main limitation of the study is that the accuracy dropped by 27% when data on the user's information was taken out. Whereas the study by Green et al. (2013) aimed to compare frequency and style-based features for Twitter author identification, Layton et al. (2010) used the feature of Bag-Of- Words and style markers. The number of users was 12; the samples from each user was 120-900 tweets; the sample size was 140 characters maximum; the number of features were hundreds for bag- of-words, and there were 86 style markers; the classifier type was SVM, and they achieved 40.5%. This indicates that the type of features selected plays a significant role in increasing and decreasing the accuracy of identification, since the features of style markers may be suitable for identifying authors in small sets of samples, but it is possibly weak

when the number of authors increases. In addition, even if Layton et al. (2010) used SVM as a classifier, their study would still be superior to Green et al. (2013) because they used the n-gram feature on Twitter. Moreover, the number of authors in the study by Layton et al. (2010) is four times larger compared to the number of users in the study by Green et al. (2013), which shows that n-gram can be more effective for short text, especially on the Twitter platform, because it deals with characteristics and words that are mostly lexicon features in text. Considering that Layton et al 's accuracy dropped by 27% when data on user information was taken out, this might be because the SCAP algorithms that they used are compatible only with certain and specified users, and so a particular user's network of communication may be necessary for determining authorship on the Twitter platform, such as following, retweeting, and replying.

On the other hand, in the study by Allison et al. (2008), the aim was to discover the authorship of emails, and they used the features word frequency, 2-grams, 3-grams and stem words. The samples for each user was nine, and the classifiers used were multimodal, hierarchical and SVM. The two classifiers, multimodal and hierarchical, are probabilistic, in that they derive an explicit estimate for the probability that a new document is appropriate for each of the likely classes. The number of features has not been specified, although the sample size was 75. Their main limitation is that complex linguistic features do not allow for successful discrimination. Their accuracy rates were 78.46% for multimodal, 87.05% for hierarchical and 86.74% for SVM. Unigram features seemed to outperform to bigrams, and trigrams, as long as there is doubt regarding certain stylistic texts; in addition, since they are captured by the longer n-grams and can contain more characteristics, this may contribute to the process of getting closer to identification. This can be compared to the studies mentioned above that used the Email

platform, as Allison et al. (2008) and Corney et al. (2002) both utilised SVM as the classifier. The study by Allison et al achieved 86.74%, while Corney et al. (2002) achieved 70.2%, bearing in mind that the message length of Allison et al. (2008) was 75 words, and Corney et al. (2002) was between 50 and 200 words. Therefore, this indicates that there are difficulties in determining the best features to be used in stylometry, the best size of Email message, and the most appropriate number of users of the Email platform, for the identification process. This suggests a need to find out more about the optimisation of these features, because there is no solid basis for relying on them during an investigation or exploration.

Lexical and syntactic features have been used for short texts, and the unique structural characteristics of messaging systems can facilitate authorship identification by using these structures and can provide important evidence which may lead to the identification of the author. For instance, words at the beginning of sentences, greetings, signatures, quotes and links, could contain important information and details that lead to understanding more about the author. However, in the case of Twitter, this might be ineffective since Twitter users tend to write informally and perhaps randomly and are restricted only by the number of words allowed. With regard to identifying and classifiying data, machine learning tools for short text have played an important role. Moreover,the machine (SVM) has often outperformed other classification methods, including: Naïve Bayes, Neural Networks, k-Nearest Neighbors, and C4.5 Decision Tree, since it can handle large-scale classification. Furthermore, the combining of two or more features, such as lexical with syntactic has been applied in many studies, as well as trying to reduce the size of the word length, with the smallest size achieved by Layton et al. (2010).

In terms of verification on most messaging platforms, broadly speaking, the majority of previous studies have focused on one platform for author verification, with different and unclear mechanisms, and with limited datasets. Little research has been found on stylometry across many of the platforms. There is also a lack of analysis of the underling feature vectors that are most appropriate for users within and across platforms.

Several key features and approaches have been discussed thus far in this section and it is clear that the level of usefulness of different features depends on the individual platform. In addition, structural features seem to increase the rate of accuracy for the Email platform and forum posts, while the most commonly used stylometric features for author identification are lexical and syntactic. Furthermore, the longer the text is, the easier it is to compute stylometric features It should also be considered that a large enough range of samples is required to ensure the accuracy of results, although SVM is useful where there is a small amount of data and it has outperformed the Bayesian classifier. Moreover, the SVM is more effective than K-NN, as it facilitates identifying the author (of books), and the SVM classifier has also outperformed Decision Tree. Overall, SVM has outperformed other classification methods because it can handle large-scale classification.

This section has explored lexical and syntactic features and has shown that these features combined can achieve high accuracy in identifying the authors of long documents due to the range of style and vocabulary. High accuracy can also be achieved if long texts contain function words such as “in”, “or”, “at”, as combining them makes it possible to achieve higher accuracy. In addition, the higher the number of samples from each suspect, the higher the recognition rate. Another factor is feature style markers, as these are useful for identifying authors from

small sets of samples, although this is less so when the number of authors increases.

It has been found that some SCAP algorithms may be compatible only with certain users Even so, unigram features can be captured by the longer n-grams and can contain more characteristics, which may increase the possibility of identification. With regard to stylometry, it is difficult to determine the best features for use in stylometry for the identification process and there is no solid basis for relying on them during an investigation. It is also significant that most studies have focused on one platform for author verification, with limited datasets and varied mechanisms. Moreover, there is a shortage of research on stylometry across the various platforms, and there is a lack of analysis of the underling feature vectors that would be most appropriate for users within and across platforms.

### User Linking

To connect authors on multiple platforms, the technique “user linking” is a quite new approach and there have been few studies, although it has been reviewed and discussed previously in the literature review. The first work on user linking was conducted by Zafarani et al. (2009), who attempted to connect users across multiple websites. Two methods were suggested: the URL of a user profile page, which contains the corresponding user’s name, and the natural user’s profile which contains another community’s username. Liu et al. (2014) attempted to build a behaviour similarity model and a structure information model, and they used multi-objective optimisation with missing information to identify linkages across social networks. Afroz et al. (2014) attempted to link users that have multiple accounts within the same forum or blog-based site; linking was based on artificially created accounts of the same user. Almishari et al. (2014 a) attempted

to link Twitter accounts based on lexical features. However, there is a weakness in all these studies for the following reasons:

1- The process of linking users depends mainly on the size and the amount of content of the text message, because often the goal of social networking sites depends heavily on messaging and text messages, even with sites that offer video and image services. For example, the comments on YouTube videos are text, Snapchat provides a text messaging service, most social networking sites, Email and Text message all provide text, thus optimisation of text using the number of words for each platform, and with multiple platforms, was not addressed in any of the above mentioned studies. In addition, most of the previous studies have introduced other factors such as name, age, sex and images, which are all reflected in the knowledge of the nature of the user's linguistic approach on these platforms. The nature of the user's linguistic approach is the most important element for understanding the nature of the writing style of the particular person, because most modern platforms provide a writing service. In addition, these have not been addressed in most previous studies of online platforms.

2- The nature of stylometric features for all platforms needs be adapted to each other and optimised using stylometric feature types. For example, the features of stylometry associated with Facebook, Twitter, or the extent of correlation features with each other; alternatively, this also makes it adaptable to the volume of text messages received in order to carry out the verification process, and this is also not available in the above studies. According to Goga et al. (2015), Jain et al. (2015) used attributes without analysing their features and their limits to match profiles in practice, therefore they used attributes with low

availability which can only match a small portion of profiles across a small number of social networks, and is likely to give many false matches in practice.

As presented in this chapter, several methods, complex techniques and uncertain systems have been proposed for solving the problem of author identification/verification on single platforms. However, it is still not clear what volume of messages is necessary for reliable and confident verification/identification, or how to approach solving the problem of author verification across platforms and whether there is the potential for stylometric features to be unified to deal with multiple messaging systems. This is because most platforms differ from each other, whether technically or linguistically, and there are several differences in most multiple platforms from different aspects, for example, for modality, Twitter is a public platform in nature, while Text message is mostly used for exchanging private text messages.

Furthermore, in terms of word length, the typical text message size for the number of characters on Text message is 160 characters, while forTwitter it is 140 or 280 characters. In addition, a user can connect to their Twitter account directly and does not need a SIM or phone to create his/her account. The other difference is that the default platform for Twitter/Facebook/Email is Internet-based, which differs from Text message. This makes finding a technique that is suitable for a range of systems highly problematic.

The most significant aspect, linguistically, is that the posts/tweets are not necessarily restricted by caution or fear of people, as with Text messages, since even though they are public, users can easily post and tweet and can hide themselves without any cost. While for SMS text messages, they must have or buy a SIM to ensure anonymity, and use caution because SMS text messages

must be addressed to a specific user, so he/she cannot deny it in most cases because he/she is the only one who sends the message to a person or specific group of people. Therefore, the writing style is less important, or in other words, there may be a lack of interest in the style of writing. The language used in Twitter and Facebook messages tends to be less formal, resulting in more misspellings and abbreviations, while Text message and Email often involve official or formal language. Whereas SMS may be seen as a one-one platform, as users message between each other, usually on a personal level.

Twitter is a public platform which involves one-many relationships, as the user is posting publicly to their many followers. This results in a different set of problems regarding illegal or unethical uses of the platforms. Whereas the danger of SMS could be that the user may not know the person messaging them, for example they could be using a fake personal information, on Twitter, issues such as encouraging hate crimes and defamation of character are more likely.

This can be seen in the case of Musk versus Unsworth, when during a television interview, Unsworth (a caver and rescuer) accused Elon Musk of a publicity stunt regarding his idea of using a pod to rescue a group of Thai boys stranded in an underground cave; Musk responded to the criticism by calling Unsworth a “pedo guy” on Twitter. Unsworth subsequently attempted to sue Musk, but Musk won the case as he argued that his comment “pedo guy” did not mention Unsworth’s name, and therefore did not constitute defamation (Mac., 2020).

This highlights the complexities around social media platforms and the need for user to be cautious about what they say, despite Musk being found not guilty, he still faced a barrage of criticism for his comments (Mac., 2020). Furthermore, had the comments been made on a private platform, the consequences would have

been far less, and the number of followers an individual has also has an impact on the seriousness of such issues.

This is why some platforms need more optimisation in order to avoid the impact of the error rate because it is likely to be relatively large, and because other platforms involve formal characters, especially when the text message is small, and this is the main reason why language platforms differ from each other.

This section has highlighted the various issues that need to be addressed, and that there are several main outstanding challenges. It is necessary to discover the optimum length of entries for each platform to support the identification of individuals. In addition, SVM should be further researched to confirm whether it is the best classifier for a small volume of data. Further exploratory research is also required to assess the usefulness of various combinations of features, in particular, syntactic features, and lexical and stylometric features together; furthermore, the impact of machine learning and the accuracy of classifiers such as Baysian, Decision Tree and SVM must be considered. Moreover, the review of the literature has revealed that it is necessary to assess both the effect of increasing the number of samples for each suspect and the impact on recognition rates for short text messages, as well as identifying the optimal number of features (which may be hundreds for bag- of-words) for identification; bearing in mind that a further comparison of unigram features with bigrams and trigrams should support the identification process. The best features for stylometry, optimum size of Email messages, and most appropriate number of users also requires further investigation. Conducting across platform research would support the analysis of stylometry and feature vectors within and across platforms for author verification, and exploring user linking across social networking sites for those with multiple accounts requires further exploration as this will support

author identification and should shed light on the different stylometric features on different platforms.

## 3.3 Conclusion

This chapter has analysed the literature on verifying authors, starting with a description of the long history of stylometry. Moreover, the key studies from the literature have been drawn on by conducting a systematic search of relevant databases to discover the most appropriate literature. Therefore, the capability of stylometric features on different modern messaging systems has been analysed, along with assessing the reliability of methods for verifying both short and long text messages. The importance of the features of stylometry has been highlighted, especially as modern messaging platforms are text-based and there is a need for adaptability in the volume of text in order to carry out the verification process; however, this has not been addressed in the studies discussed. In addition, with regard to users' profiles being transferable between systems and whether there are common stylometric user characteristics, this also has yet to be researched. It has been shown that it is important to use attributes with high availability in order to avoid false matches, but the optimum volume of messages for reliable author identification remains unclear. In addition, approaches vary across platforms as they are used for different reasons and therefore involve different writing styles, for example Email messages are more formal compared to Facebook posts. Hence, author verification on modern messaging systems is complex and requires further research in order to improve verification rates and find individuals engaging in activities such as trolling and other criminal behaviour.

# Chapter Four: Research Methodology

## 3.4 Introduction

Prior research has clearly shown that there is some ability, or some degree of performance, that can be achieved in authorship verification on various platforms. However, questions remain due to issues around dataset size and relative performance; for example, whether the linguistic characteristics for writing tweets are different from writing Emails or text messages. No previous studies have been found that facilitate a direct comparison between the performance of authorship verification methods across messaging platforms. It is important to examine relative performance across platforms for the following reasons: Firstly, to understand which of these modern platforms is more reliable and shows better performance concerning sources of data for authorship attribution. For example, Facebook is far better at providing reliable sources of data for author attribution, whereas Twitter is perhaps not so reliable; therefore, it is necessary to compare data from one user across modern messaging systems' platforms. Secondly, an analysis of feature sets will assist in understanding the role of linguistic characteristics between platforms and should result in discovering what stylometric features of short texts are shared between multiple messaging systems. This is because of the current incompatibility in profiles across multiple platforms. This should support the creation of appropriate and sufficient information to provide a reference template and perform verification, as current systems do not permit a direct comparison across systems. For example, a suspect may have a legitimate and benign Facebook profile through which he communicates with friends, and meanwhile, he may engage in criminal activity on Twitter. Furthermore, at present, there are limitations from using Twitter-based features within the verified Facebook linguistic profile of a user. The problem

presented in this research is unique and different to previous research in that it has focused on exploring the relative performance of authorship verification across messaging platforms, resulting in greater understanding of the nature of stylometric features of a suspect that can inform working across platforms.

This research has investigated the relative performance of authorship verification across platforms and has explored the identity of an author according to short volumes of text. By comparing users' stylometry performance across messaging platforms, stylometric features can be identified that are platform-dependent and independent, including for forensic investigation. Given sufficient platform-independent features, it may be possible in the future to develop a platform-dependent stylometry profile, which would allow for creating a unified reference template to facilitate platform biometric independent author verification that could be used for linguistic forensic investigation. Therefore, this research could provide a starting point for future research in that direction.

In addition, to aid the recognition process, a better understanding of the volume of information is required. For example, it would be useful for the analyst to know with what level of confidence an author verification decision is made, and to what degree this is dependent on the length and characteristics of the message. Therefore, this research has also investigated the nature and volume of text required to support the underlying recognition.

Hence, this research will address the following four research questions:

- RQ1: What is the relative performance for the population and single users on single platforms and across platforms, including relative performance across multiplatforms for the same user?

- RQ2: What feature vector and composition are most viable across messaging platforms and what is the impact on performance?
- RQ3: To what extent is it possible to identify and derive platform dependent and independent stylometric features with a view to enabling platform-dependent author verification?
- RQ4: What is the minimum set of information that would be required to provide reliable verification of an author? (This would measure and characterise the limitations with respect to message length and composition to provide reliable author verification decisions).

This chapter presents the methods used to collect data, the approach used to prepare the data to support the aforementioned experiments and messaging platform samples, and the software that was employed. This is followed by discussing feature selection, the pre-processing of data, feature extraction, selecting influencing features, splitting the data features tested, and finally, classification modelling.

## 3.5 Research Methodology

Deciding on the selection of a research methodology is an important element and a fundamental aspect of any research, as this should lead to finding the correct answers to the research questions precisely and accurately; on the other hand, inadequate selection would lead to inaccurate answers to the research questions. The main types of research methodology are: Quantitative, Qualitative, Deductive, Pragmatic (mixed approach), and Advocacy/participatory approaches (Creswell et al., 2017).

*Quantitative research* is an approach used for exploring and understanding the meaning individuals or groups attribute to a social or human problem. The

research process involves emerging questions and procedures, with data often collected in the participants' setting, and data analysis inductively built up from particulars to form themes; in addition, the final written report has a flexible structure (DeLeeuw., 2018). It is a process that includes collecting, modifying and converting data into numerical values to form statistical assumptions. For example, online surveys, mail surveys, paper questionnaires, face-to-face questions, and telephone interviews. Objectivity and data sensitivity are significant in quantitative research, therefore investigators must take good care to avoid their own perspective, behaviour or attitude affecting the results. It is an approach used in the examination of objective theories by investigating the relationship between variables; often, these variables are related to the positivist/post positivist pattern, and can be measured using instruments, and the numbered data can be analysed using statistical procedures (Silverman, 2016).

*Qualitative research* is investigative research and it is used to reach an understanding of attitudes, opinions, behaviours, motivations and perspectives using a small sample from a larger population (Creswell et al., 2017). Basically, it investigates and attempts to uncover hidden conceptions and consequences of human behaviour. Most researchers using this approach are concerned with gaining a rich and complex understanding of specific occurrences in society, rather than gaining information that can be generalised to larger groups. Examples of qualitative methods are individual interviews, focus group interviews, and observations.

*Deductive*, is concerned with developing a hypothesis based on existing theory, and then designing a research strategy to test the hypothesis. It explores the relationship or link that seems to be implied by a particular theory or case example, and it might be true in many cases. It may therefore test to see if this relationship

or link is relevant to more general circumstances (Jonathan, 2010; Gulati, 2009). Deduction begins with an expected pattern that is tested against observations, and seeks to find a pattern within them (Cramer-Petersen et al., 2019). The deductive approach has the following advantages: possibility to explain causal relationships between concepts and variables, the option to measure concepts quantitatively, and possibility to generalise research findings to a certain extent.

*Pragmatic* (mixed approach) is an approach to inquiry that involves collecting both quantitative and qualitative data, thereby integrating the two procedures of data collection. It uses different ways of interpreting the data since no single point of view can provide a complete understanding of a research problem, and it gives the entire picture rather than using either approach alone (McBride et al., 2019). In addition, a mixed approach allows for triangulation of the data, which can give more weight to the research findings (Creswell et al. 2017).

*Advocacy/participatory*, sometimes named emancipatory, is where researchers adopt "an advocacy/participatory approach feel that the approaches to research described so far do not respond to the needs or situation of people from marginalised or vulnerable groups. As they aim to bring about positive change in the lives of the research subjects, their approach is sometimes described as emancipatory" (Shirish, 2013).

Consequently, a qualitative approach is the most appropriate method for this research because it can be used in investigating and understanding of attitudes, opinions, behaviours, motivations and perspectives using a small sample from a larger population. Basically, it investigates and attempts to uncover hidden conceptions and consequences of human behaviour. Furthermore, as the sample size is a small sample from a larger population (the empirical basis for evaluating

the research question was to engage 50 participants), a qualitative approach is most suitable. In order to conduct the aforementioned experiments effectively, various types of modern text messaging corpora have been collected from users' messaging systems (Twitter, Text message, Facebook, Email) to explore the significant features learned from their writing styles. Users' messaging samples highlight the linguistic feature differences, and the unified experiments have included combining samples with each other from on different platforms to verify the user; whereas, portable experiments have been used for feature sets to form a superset that could be used to test a text sample against another sample from a different platform; this is referred to as cross-domain datasets. Finally, the text length sample of the users was investigated to assess the impact on the reliability of author verification decisions.

This allowed the following related aspects to be explored:

- Understanding the performance of single messaging systems and investigating the impact that feature length and the composition of the feature vector have on performance.
- Investigating what commonalities and differences exist within the feature set, as well as what commonalities and differences exist within the feature set across platforms.
- Understanding the performance of unifying and portability for messaging system verification using multiple text message samples, and how this performance compares across platforms.
- Understanding what the minimum set of information is that would be required to provide reliable verification of an author.

To illustrate the desired investigation further, Figure 0-1 below illustrates the methodology used in the research.

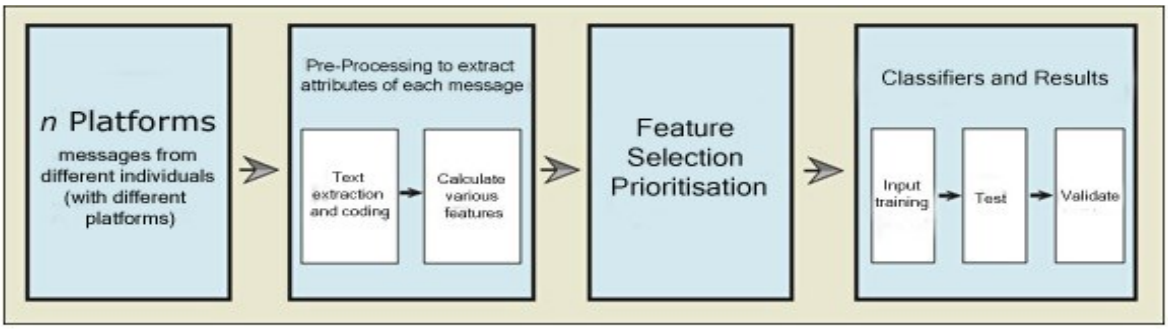



Figure 0-1: Research Methodology

When looking to investigate the extent to which feature vectors could be used across platforms, the research sought to take two approaches: The first approach focused on unified features, which involved combining the feature sets of different platforms and then prioritising the critical features. The second approach involved examining a particular subset of features identified that were common across the platforms for portability. Figure 0-2 reflects the feature spaces of these two problems to highlight which feature was being used.

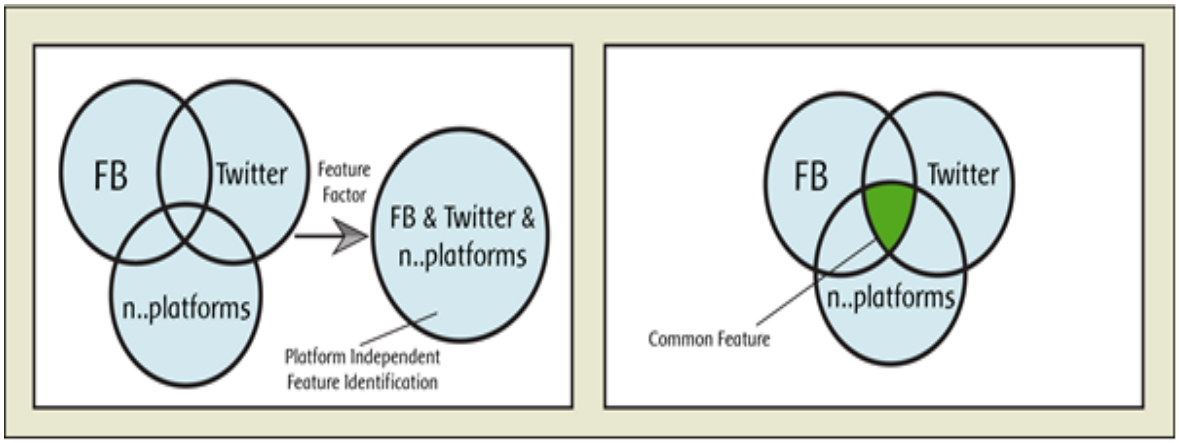



Figure 0-2: Feature Set Abstraction

In this research, five main steps were used in the research method: data collection, feature vector extraction, feature importance analysis, train/test

splitting ratio, and classification modelling, which are discussed in the following sections.

## 3.6 Data Collection

This section presents the scientific and methodical approach to the collection of real data contained in text messages for the core messaging systems of Text messages and Email, and the social messaging networks of Twitter and Facebook. This stage of the research was the most challenging since many participants consider their text messages to be sensitive and to contain highly private content, making it difficult to negotiable a way for them to share the content of their messages for the purposes of the research. In collecting the data for this research, consideration had to be given to the privacy of the information because, clearly, the users would not be willing to participate in a study that required them to hand over the entirety of their Text messages, Email messages, and so on. Therefore, the solution to the problem was to create a data collection process that enabled the computation of the features on the client's machine. This allowed two things: firstly, it ensured that no information or private details were taken from the users' platforms; secondly, ethical approval had to be obtained. This meant that unlike most studies into biometrics, it was necessary to identify all possible features that needed to be collected ahead of time.

The goal was to collect as many text messages as possible and as many users as possible, while the historical data for this research targeted authors who have more than two messaging systems. The authors should have had a variety of messaging systems (Text messages, Email, Twitter, Facebook), with a minimum of two platforms in order to be targeted and be of consenting (age 18 years +). Whilst it would have been useful to insist on all four platforms, there was a concern that this would impact on the ability to recruit a sufficient number of

participants. The methodology used in the data collection, a description of feature selection and features extraction, and data prepossessing, are presented in the following sections.

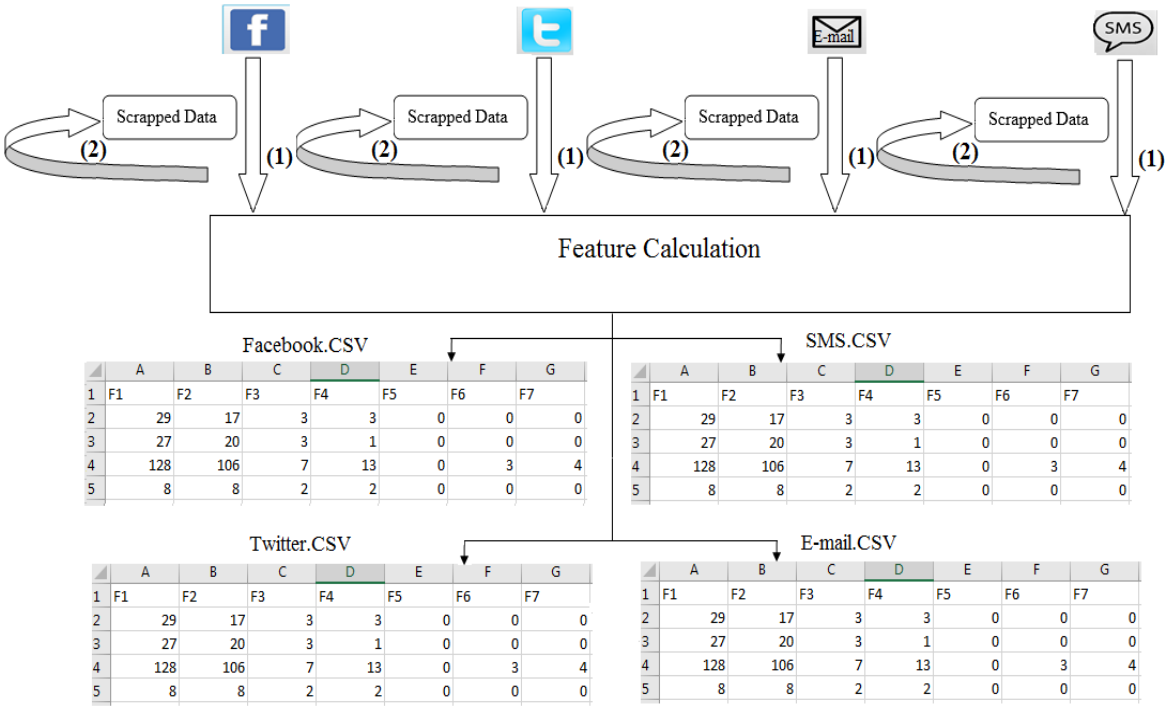
### 3.6.1 Messaging Data Collection

In general, the research has sought to explore the author verification techniques that can be used to verify individuals based on the composition of their messages. Having stated that, such approaches need to be developed on a per-system basis (i.e. the profile used to verify an individual from Text messages would be different to Email), to understand and investigate how much text is required, as well as to explore to what degree a unified single profile can be used across messaging systems.

Since it was not possible to see the users' plain text messages on the messaging systems, stylometric features were designed before the data collection to ensure that the software and application were working as required. In addition, ethical approval was acquired from the university's Research Ethics Committee before proceeding (see Appendix D).

The participants were asked to sit in front of a computer machine and perform a set of logins to provide access to (up to 4) messaging systems: Facebook, Twitter, Email and Text message. A tool was used to extract their text messages (the tools for exporting participants' text messages have been illustrated in section 4.3.4). In this manner, the highly private messages have not been stored or used directly, and the researcher simply removed the text messages and the extracted feature sets (note that these do not contain any information that could be used to recreate the original message) and then the necessary features required were calculated (the stylometric feature design is illustrated in the next section - 4.3.2). Figure 0-3

below illustrates the data collection methodology and the process and the resulting data that has been captured.




| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | F1 | F2 | F3 | F4 | F5 | F6 | F7 |
| 2 | 29 | 17 | 3 | 3 | 0 | 0 | 0 |
| 3 | 27 | 20 | 3 | 1 | 0 | 0 | 0 |
| 4 | 128 | 106 | 7 | 13 | 0 | 3 | 4 |
| 5 | 8 | 8 | 2 | 2 | 0 | 0 | 0 |

Figure 0-3: Data collection methodology

Any scrapped data (containing the participants' messages) has only been stored temporally for use in the feature calculation application.The output file is a set of features, as shown in Figure 0-3; the file contains no private information, and it is not possible to go from these features back to the original content, as it is a one way process. The scrapped data was deleted and removed from the hard drive during the session which the participant was present at and has not been taken forward at all. When each user had finished the session, there was no way the researcher could to go back to their account; there was no way to access their data, no relationship, and no information about the message itself. The procedure of the time extraction took about 15-30 minutes for each platform. The participants were thanked for their participation and told how they could obtain further information about the research. They were also told how to contact the research team if they needed to later on.

As the experiment was carried out in the Centre for Security, Communications and Network Research (CSCAN) at the University of Plymouth, with the dedicated experiment time in mind, and the main targeted sample was PhD researchers at CSCAN. Also, other University of Plymouth postgraduate and undergraduate students were invited to participate. The participants were not asked about their demographic information such as name, gender, age and ethnic background. They were only asked if they are a student at Plymouth University because the study has focused on the worst case scenario where the suspects' demographic information is not available to the investigator. This is important because during the stage of an investigation, a crime investigator may not have any clue about the potential suspects's demographic information - only the area and the place where the spatial message has come from for the given tweets. The research has attempted to gain a deeper insight into the writing styles of the given text messages written by the same authors. In order to facilitate a meaningful analysis, the total number of subjects targeted was 50 as a minimum, and a total of 50 participants were gathered as the final outcome, which is considered a sufficient baseline according to other previous research that has been conducted using approximately similar sample sizes (Li et al., 2014; Zheng et al., 2006).

### 3.6.2 Feature Selection

Having explained that since privacy factors made it impossible to see the users' plain text messages on the messaging systems, stylometric feature calculations were designed and selected before the participants provided their text message data. Designing and building the stylometric features selection and data processing before collecting the data and all necessary procedures, ensured that no text messages showed any plain text from the users' platforms. It has been shown that since privacy factors can make it impossible to view the users' plain

text messages on the messaging systems; therefore, stylometric feature calculations were conducted prior to the participants providing their text message data. The designing and building of the stylometric feature selection and data processing was carried out before collecting the data and performing the procedures, which ensured that none of the text messages revealed any plain text from the users' platforms.

### 3.6.3 Stylometric Features

As mentioned previously, the goal was to collect as many text messages as possible, therefore, a portion of the 227 stylometric features was selected from a subset of features from Zheng's research (Zheng et al., 2006 ) and Li's research (Li et al., 2014), to include character-based and word-based features, syntactic, structure and social specific features. The main reason for this is that their stylometric features have achieved good performance with online messaging systems and social media accounts. The reason for considering stylometric features selecting from a subset of features from Zheng's research is because: firstly, they studied the authorship of online messages and dealt with texts from online messages including Emails, newsgroups, and chat rooms. Moreover, their study of messages from newsgroups or chat rooms may have similar characteristics and short texts to social messaging systems such as Twitter, Text message and Facebook. Furthermore, similar to chat rooms or newsgroups, Twitter, Facebook, and Text message provide a sociable and casual environment for users to share information and communicate with each other. Secondly, the average word count for Zheng's data was 169 words, which is relatively short compared to other research studies (Hussain et al., 2014; Monaco et al., 2013; Li et al., 2014; Iqbal et al., 2010; Pavelec et al., 2009; Stamatatos, 2007). Thirdly, they achieved significant results: 97.69% accuracy rate with SVM.

Moreover, in order to increase the number of social linguistic features as much as possible, a subset of features from Li's (2014) research was considered due to their stylometric social features, and because: Firstly, they used posts from the Facebook platform. Secondly, their performance was 79.6% as an accuracy rate, approximately EER ≈ 20%, which may be considered somewhat reasonable since their performance was high.

Moreover, in order to increase the number of linguistic features for the Text message platform as much as possible, a subset of features from the research by Saevanee et al., (2011) has been considered. This is because of the stylometric features they explored and because they have described the Text message platform, as well as achieving an EER of 24%.

Moreover, 48 additional features popularly used on social media such as emotional icons have been included (it is the first time that these emoticon features have been used and tested in this way). A total of 275 features, including 227 stylometric, and 48 social network specific features with emoticon features, were extracted. A comparison of selecting stylometric features was conducted, and Table 0-1below shows the stylometric features used across platforms in the research (Zheng et al., 2006; Li et al., 2014; Saevanee et al., 2011); in addition, Table 0-2 presents an overview of the feature groups selected within each platform, and a complete listing can be found in Appendix B.

Table 0-1: A comparison of stylometric features

| Studies | Platforms | Feature type | | | | |
|---|---|---|---|---|---|---|
| | | Lexical | Syntactic | Structure | Social specific | Emotional icons |
| (Zheng et al., 2006 ) | **Email**, newsgroups, chat rooms, **Twitter** and online messages. | √ | √ | √ | | |
| (Li et al., 2014) | **Facebook** | √ | √ | √ | √ | |
| (Saevanee et al., 2011) | **SMS** | √ | √ | √ | | |
| **This research** | Twitter, Facebook, Email, SMS | √ | √ | √ | √ | √ |

Table 0-2: A summary of stylometric features

| Feature Type | Features | Description |
|---|---|---|
| Lexical features | Char based (F1-50) | Character-based features (features 1-50), which count the frequency of specific characters such as number of Alphabetic, characters, Special characters and uppercase that will be tested. |
| | Word based (F210-227) | Word based features (features 210-227), such as counting the frequency of long words or short words will be tested. |
| Syntactic features | Punctuations (F51-58) | A set of punctuation listed from (features 51-58) will be tested. |
| | Function words (F59-208) | A set of function words listed from (features 59-208) will be tested. |
| Structural features | No of sentences (F209) | (Feature 209), which shows the number of sentences will be tested. F213 and F214 can calculate sentences feature and can be categorized for structural feature or word based lexical feature. |
| Social specific network features | Social network specific and emotions (228-275) | Such as emoji, and emotional icons and missing proper punctuation listed from features. |

A program was developed and some of the software was purchased to assist with extraction features, and these are discussed in the following sections.

### 3.6.4 Exporting Text Messages

This section explains, in detail, the process used to export participants' sample text messages and the steps followed. It also contains a detailed explanation of the procedures for exporting data from each platform, along with presenting the program that was developed to extract the stylometric features from the sample of users. The software utilised to export the user samples for each platform is presented below.

Table 0-3: Software of data collection used

| Steps | SMS | Twitter | Facebook | Email |
|---|---|---|---|---|
| (1) Data Source Connection | Jihosoft Phone. Available online: https://www.jihosoft.com/mobile/phone-transfer.html | Data Twitter API. Available online: https://developer.twitter.com/en/docs.html | Data Facebook API. Available online: https://developers.facebook.com/docs/apis-and-sdks/ | Export Outlook Emails. Available online: https://outlook.live.com |
| (2) Additional Software | JSON to CSV. Available online: https://json-csv.com | - | - | ReliefJet Essentials. Available online: https://www.reliefjet.com |
| (3) Feature Processing | Feature Extraction Passer | | | |

An automated feature extraction program (see Appendix D) was developed by using NetBeans for feature extraction, which is an Integrated Development Environment for Java. (the process of calculating of features will be described later).

### 3.6.5 Data Pre-Processing

Each of the messaging platforms required a process to be developed to parse the relevant messaging data, ensuring only relevant data was parsed. For example, for Email, it was important to ensure the user's Emails were parsed and not the replies to Emails that are often appended. Given the nature of the sensitivity of the data, it was critical that this parsing did not simply extract the messages but automatically performed feature extraction and was run on the participant's computer. Therefore, the raw data needed to be extracted and converted for use within the feature vector extraction utility. Each platform required a bespoke solution for the pre-processing procedures, which are as follows:

- Email: Outlook Emails from the folders "sent" and "sent items" within each user's Email account were selected; all duplicated Emails and signatures were removed. The preprocessing was carried out automatically using scripts and was not done manually; each Email was parsed to extract the body of the message and remove received texts when they existed. All Emails that contained titles, tables and web addresses were removed.
- Text message: Software called *Jihosoft Phone Transfer* was used to export the data from the participants, as described earlier in Table 0-3, and a feature extractor program was used to parse each SMS text to extract the body of the message and remove received texts if they existed. All Text messages that contained numbers, titles, tables and addresses were removed.
- Twitter: A data crawl from the Twitter API was used to return a list of all tweets of a given participant. All duplicated tweets and Re-Tweet (RT) tweets were removed. All tweets that contained pictures, tables and web addresses were removed.

- Facebook: A graph Facebook API was used to return a list of all posts of a given user. All posts that contained numbers, titles, tables and web addresses were removed.

A number of scripts have been developed and generated in order to perform a variety of functions to implement the pre-processing for each platform (see Appendix E).

## 3.7 Feature Vector Extraction

During the data collection process, samples of users' Twitter, Email, Facebook and Text messages were passed through a procedure to extract all of the features from each user's messages. An automated feature extraction program (see Appendix D) was developed by using NetBeans for feature extraction, which is an Integrated Development Environment for Java. NetBeans IDE supports the development of all Java application types, and it is common and well-accepted software commonly used in scientific and developer communities for languages.

The program calculates the features during the data collection, and the output contains only the calculation of numbers to ensure and maintain the privacy of users. It works by reading the individual input files that have been extracted from each Twitter, Text message, Facebook and Email message. The input of the program was a text file that contained all messages made by an individual. Each line represented an SMS text/post /Tweet/Email of that individual. The steps for extracting the features executed by the newly developed program are:

1- Read a line (representing a post/Tweets/ Text message/Email).
2- For feature one to feature 275, measure each feature from the post/Tweet/Text message/Email. Each feature and its measurement have been stored as a feature vector (name and value pair).

3- All 275 name and value pairs were written to an output file for each message.
4- Steps one to three were repeated to read the next line if there were more text messages.

The output of the developed program is a text file (Microsoft Excel Comma Separated Values File (csv)) with the same number of lines as the input file. For each Tweet/post/Text message/Email message, all 275 features were measured. Each feature and the value have been represented as a feature vector, ultimately a name and value pair *(Fn: Y) where* Fn is the feature n and Y is the value of measuring feature n from that Tweet/post/Text message/Email message, while n ranged from one to 275. Figure 0-4 shows an example output of a developed program.

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | F9 | F10 | F11 | F12 | F13 | F14 | F15 |
| 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 19 | 5 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 2 |
| 4 | 20 | 14 | 2 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 2 | 1 | 0 | 0 | 2 |
| 5 | 8 | 5 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 2 |
| 6 | 120 | 102 | 7 | 9 | 1 | 3 | 6 | 9 | 3 | 0 | 7 | 11 | 1 | 0 | 5 |
| 7 | 99 | 72 | 5 | 3 | 0 | 1 | 0 | 10 | 1 | 0 | 6 | 5 | 0 | 1 | 2 |
| 8 | 5 | 5 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 2 |

Figure 0-4: Output of the developed program

In terms of privacy, the only thing that appears when collecting data is the interface of the automated feature extraction software, as shown in Figure 0-5 below.

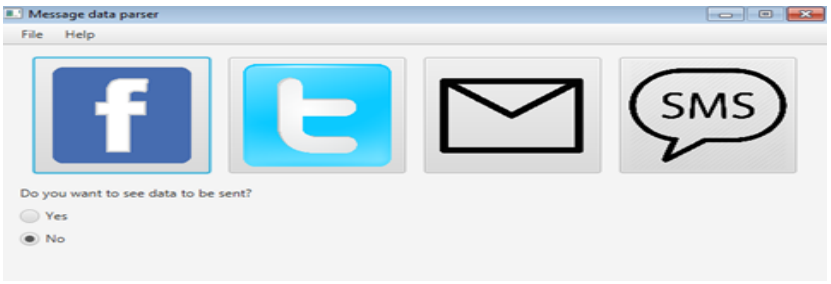



Figure 0-5: A screenshot of the interface of an automated feature extraction software

A descriptive example of how a text message was converted into feature vectors is represented in Figure 0-6 below. F1: Refers to the list of features (see Appendix B); the first feature vector (732) indicates that the value for Feature 1 (the number of characters) was 732. The second feature vector for Feature 2 (number of alphabets) was 701. The rest of the feature vectors were also measured. There were 275 feature vectors in total for each Tweet/post/ Text message/Email message. The output file had the same number of lines as the input file, indicating that the features were extracted from each sample.

| | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | user | sample | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | F9 | F10 |
| 2 | 1 | 1 | 732 | 701 | 22 | 48 | 6 | 28 | 30 | 104 | 13 | 7 |
| 3 | 1 | 2 | 377 | 356 | 13 | 25 | 2 | 5 | 12 | 43 | 7 | 8 |
| 4 | 1 | 3 | 434 | 407 | 16 | 33 | 5 | 11 | 16 | 47 | 8 | 9 |
| 5 | 1 | 4 | 9 | 6 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| 6 | 1 | 6 | 186 | 174 | 7 | 18 | 0 | 10 | 9 | 29 | 3 | 2 |
| 7 | 1 | 7 | 116 | 106 | 5 | 10 | 0 | 7 | 6 | 14 | 2 | 1 |
| 8 | 1 | 8 | 354 | 337 | 10 | 21 | 4 | 8 | 13 | 36 | 6 | 7 |
| 9 | 1 | 9 | 301 | 279 | 16 | 25 | 4 | 10 | 8 | 26 | 5 | 6 |

Figure 0-6: A text message converted into feature vectors

## 3.8 Historical Dataset Desired

In order to give an overview of the desired historical datasets and view the number of samples per user across the four modern platforms, the overall final corpora, including the total number of samples for each user, was revealed. After the data

was collected from the participants, pre-processing was performed, and the feature vectors were calculated. As a result, as shown in Table 0-4, the data encapsulated from the 50 contributors (at least two messaging platforms must be available for a participant, also at least 20 messages must be obtained from the user on one corpus) could be considered deep enough to enable a significant analysis. There has not been any previous research that has examined a real life dataset in this way (to the best of the researcher's knowledge).

Table 0-4: The Overall Final Dataset Statistics

| Description | Platforms | | | |
|---|---|---|---|---|
| | SMS | Twitter | Facebook | Email |
| Number of participants | 26 | 41 | 46 | 47 |
| Number of text messages | 106,359 | 13,617 | 4,539 | 6,540 |
| Average number of text messages per user (mean) | 4091 | 332 | 99 | 139 |
| The maximum length of text messages | 30.6 words | 35 words | 1147 words | 3712 words |
| Average length of messages per user | 10 words | 13 words | 15 words | 74 words |

Table 0-5 shows that the total number of samples were collected from each user across platforms. The total number of samples in each corpora was 13.617, 106.359, 4.539 and 6.540, for Twitter, Text message, Facebook and Email respectively.

Table 0-5: Total users for each platform

| User | Facebook | Twitter | Email | SMS | Total | #platforms |
|---|---|---|---|---|---|---|
| 1 | 71 | 583 | 83 | 19,141 | 19,878 | 4 |
| 2 | 46 | 20 | 161 | 403 | 630 | 4 |
| 3 | 27 | 599 | 72 | 37 | 735 | 4 |
| 4 | 28 | 579 | 30 | 1,718 | 2,355 | 4 |
| 5 | 189 | 584 | 386 | 852 | 2,011 | 4 |
| 6 | 90 | 595 | 21 | 6,071 | 6,777 | 4 |
| 7 | 48 | 590 | 202 | 2,687 | 3,527 | 4 |
| 8 | 95 | 270 | 49 | 1,279 | 1,693 | 4 |
| 9 | 76 | 590 | 80 | 4,729 | 5,475 | 4 |
| 10 | 68 | 146 | 51 | 3,611 | 3,876 | 4 |
| 11 | 56 | 105 | 38 | 29,710 | 29,909 | 4 |
| 12 | 139 | 46 | 314 | 207 | 706 | 4 |
| 13 | 76 | 587 | 109 | 45 | 817 | 4 |
| 14 | 117 | 594 | 39 | 5,243 | 5,993 | 4 |
| 15 | 97 | 596 | 125 | 25 | 843 | 4 |
| 16 | 106 | 106 | 43 | 523 | 778 | 4 |
| 17 | 69 | 575 | 145 | 10,596 | 11,385 | 4 |
| 18 | 71 | 26 | 34 | 909 | 1,040 | 4 |
| 19 | 132 | 591 | 165 | 0 | 888 | 3 |
| 20 | 175 | 0 | 79 | 7,512 | 7,766 | 3 |
| 21 | 189 | 151 | 24 | 0 | 364 | 3 |
| 22 | 37 | 589 | 20 | 0 | 646 | 3 |
| 23 | 142 | 176 | 20 | 0 | 338 | 3 |
| 24 | 26 | 0 | 38 | 4,499 | 4,563 | 3 |
| 25 | 216 | 0 | 26 | 548 | 790 | 3 |
| 26 | 145 | 586 | 22 | 0 | 753 | 3 |
| 27 | 131 | 0 | 120 | 27 | 278 | 3 |
| 28 | 35 | 590 | 178 | 0 | 803 | 3 |
| 29 | 51 | 62 | 129 | 0 | 242 | 3 |
| 30 | 0 | 22 | 83 | 979 | 1,084 | 3 |
| 31 | 140 | 163 | 35 | 0 | 338 | 3 |
| 32 | 195 | 98 | 774 | 0 | 1,067 | 3 |
| 33 | 34 | 184 | 28 | 0 | 246 | 3 |
| 34 | 29 | 573 | 66 | 0 | 668 | 3 |
| 35 | 208 | 87 | 1,323 | 0 | 1,618 | 3 |
| 36 | 100 | 583 | 104 | 0 | 787 | 3 |
| 37 | 145 | 564 | 28 | 0 | 737 | 3 |
| 38 | 39 | 0 | 0 | 627 | 666 | 3 |
| 39 | 23 | 120 | 0 | 4,237 | 4,380 | 3 |
| 40 | 86 | 0 | 71 | 144 | 301 | 3 |
| 41 | 97 | 578 | 214 | 0 | 889 | 3 |
| 42 | 128 | 26 | 96 | 0 | 250 | 3 |
| 43 | 200 | 211 | 53 | 0 | 464 | 3 |
| 44 | 0 | 26 | 30 | 0 | 56 | 2 |
| 45 | 0 | 20 | 23 | 0 | 43 | 2 |
| 46 | 109 | 0 | 116 | 0 | 225 | 2 |
| 47 | 72 | 0 | 310 | 0 | 382 | 2 |
| 48 | 60 | 406 | 0 | 0 | 466 | 2 |
| 49 | 0 | 20 | 74 | 0 | 94 | 2 |
| 50 | 126 | 0 | 309 | 0 | 435 | 2 |
| **Total** | **4,539** | **13,617** | **6,540** | **106,359** | | |
| Mean | 99 | 332 | 139 | 4,091 | | |
| Median | 86 | 157 | 72 | 909 | | |
| No of users | 46 | 41 | 47 | 26 | | |

## 3.9 Selecting Discriminating Features

After the feature vectors were extracted from the participants' data, selecting the most discriminative or effective of the 275 generated feature vectors for a promising author verification profile cross platform was crucial, along with prioritising the features in terms of discriminative information prior to being applied to a standard supervised training methodology. An algorithm (Ranked Features) was employed by choosing the most important feature set: Random Forest Classifier. Random Forest prioritised the feature vector as a result of experimenting with the feature vector length and performance. Random Forest algorithms (RF) was selected (Torgo, 2016; Chris, 2017) because it has the ability to assign importance to features, giving a direct indication of the weights for each feature type (Torgo, 2016; Chris, 2017), as well as facilitating finidng the most robust features that distinguish users' samples. Therefore, it is possible to identify a subset of the most important features based on their contribution to the decision being made by the algorithm. The Random Forest algorithm plays an essential role in data science and is commonly used for feature selection in a data science workflow (Torgo, 2016; Chris, 2017), mainly focusing on treating decision trees as weak learners, and randomly subsample sets of features from a specific training dataset, to improve accuracy and mitigate overfitting (Maitra et al., 2016). Only the top *n* ranked features were fitted into the classifier in order to organise them, which made it possible to identify a subset of the most important stylometric features based on their discriminative contributions. So, after extracting the platform features from the raw data, it was fed into the Random Forest classifier to find the most robust features on both a population-base and a user base. A population base is (across all users), while a user-base is (across the authorised user) in order to permit an analysis of the impact on recognition performance.

Practically, the ranked features present good feature quality, thereby highlighting the high performance, attainment and speed of the classifier's computation. Figure 0-7 illustrates the process of sub-setting the feature set.

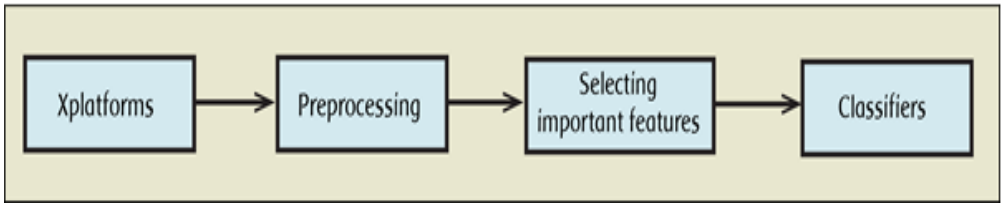



Figure 0-7: Methodology for selecting discriminative features

The ranking has been used to prioritise the features in terms of discriminative information in all experiments accodring to the two scenarios below. Based on the prior art, there are typically two approaches to feature vector composition and the analysis of the population of users within the group; in addition, more recently, the composition of feature vectors has been based on the analysis of individual users and recognsing them by the way they interact with the wider population (Clarke & Furnell, 2017).

- **Scenario 1**

- To find the most robust features in a population-base (across all users).

In order to understand the performance of the messaging systems and the recognition of population based-features, including how their performance compares to prior work, the dataset containing all users' samples was fitted into the Random Forest algorithm (RF) to identify only the most relevant features. The RF algorithm deals with this as a multi-class classification problem. Only these top ranked features were fitted into the classifier to class them based on a two-class problem in order to verify them, as shown in Figure 4-8.

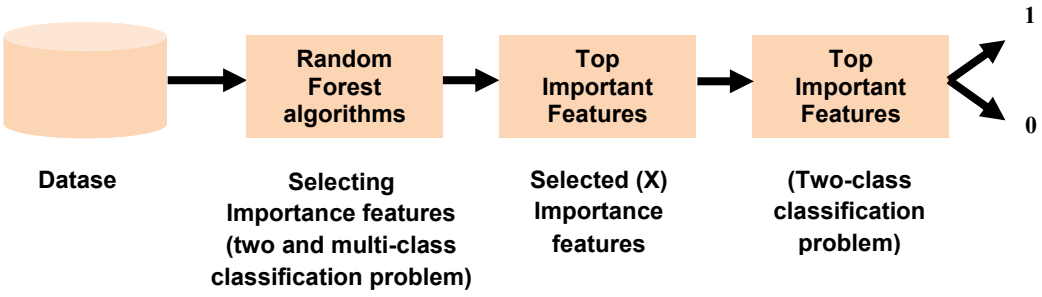



Figure 0-8: Methodology for algorithms (RF) in population and individual -based feature

➢ Scenario 2

- To find the most robust features from a user-base (across authorised users).

  To determine the performance of messaging systems' recognition for individual based features, the dataset containing all the users' samples was fitted into the RF algorithm to identify only the most relevant features. The Random Forest algorithm deals with this as a two-class classification problem. Only the top ranked features were fitted into the classifier to class them based on a two-class problem in order to verify them, as shown in Figure 0-8.

## 3.10 Dataset Handling Splitting Ratio

A key requirement in biometric design is the identification of a potential setting for classifiers and keeping only the most influential features for each individual platform, because each messaging system has unique stylometric features which are characterised by it. Furthermore, selecting the appropriate settings for the classifiers was necessary since, firstly, it helps in adjusting the discriminating features; secondly, it undoubtedly allows for reducing the noise of the classifiers by removing redundant and irrelevant features; as a result, the classification will be more accurate. The splitting data in the training/testing phase is an important aspect to train the base stylometric features, the datasets for Twitter, Text

message, Email and Facebook were split into 70/30, 60/40, 40/60, 50/50, 20/80, 10/90 for training and testing sets respectively, in order to investigate the effectiveness of ratios and divide the train/test changes on the system's performance. In addition, the reasons for selecting these settings for both classification and feature testing is because often when the training data is small, and the testing data is large, or vice versa, the parameter estimates will have greater variance, leading the performance statistics having greater variance and the results to be non-neutral. However, factors such as data splitting and model classifiers may lead to unintentional discrimination, resulting in a systematic disparate impact (Chen et al., 2018). Therefore, all possible possibilities have been addressed by minimising the trained data and increasing the data tested, and vice versa, for all three classifiers ((Support vector machine (SVM), Random Forest (RF) and Gradient Boosting (GB)) to find the best stylometric features, including checking variants between them, (why these classifiers were selected is explained in more detail in the Classification Approaches section 4.9). Once the best split was chosen (the one that achieved highest performance), and the tested feature could ultimately be determined and selected based on the best splitting data found for the final dataset performance.

## 3.11 Feature Vector Length

In addition to the composition of the feature vector being important, its length is also important because of the problem of curse dimensionality (Elkahky et al., 2015; Akkarapatty et al., 2017). It is important to ensure the length is optimal in order to achieve the best performance, and it is simply not viable to have as long a feature vector possible. Therefore, In this study, 275 stylometric features were proposed for use in authenticating users. In order to create individual user profiling, special categories such as abbreviation and emotion-based words that

a user uses in their messages were selected, as these special words may provide some useful insights into the verifying of the author. Accordingly, the following features were included for each modern corpora: lexical, syntactic, structural, short message features and social emotional features. These features have been integrated into a stylometric feature set because they may contribute towards providing discriminating features for platforms, and consequently, between authors.

In comparison, feature vector composition is the process of making changes to a feature to add newer features and make modifications to the already existing features. Each of these features is supposed to have a characteristic that is considered to be useful, intuitive, and effective. Testing more features is extremely important in order to find any common characteristics relevant to the implementation of commonality across platforms, for example lexical features, and they should be tested thoroughly at every stage of the investigation. A large number of features would place a burden on the classification, therefore the aim of testing features is to ensure selecting the best feature that works properly and meets all the intended specifications for the classification. This resulted in finding the best match from among well accepted testing features in order to validate the effectiveness of the 275 features generated to produce a promising verification technique, which is indeed an important aspect for feature vector length. Therefore, the features tested were divided into 10, 20, 30, 50, 100 and 275 in the first experiment, because it was observed during the testing of features that using more than ten new features made the performance changes more noticeable. Furthermore, since the lowest user verification features are one of the goals, and in order not to cause inconvenience and increase the complexity of the classifiers, the first top 100 features were utilised. Moreover, to increase the

test results and to discover the effect of all features, they were tested together. Since this led to good results, it was also employed in the second experiment. Since the stylometric features depend on the application used (Belvisi et al., 2020), and since this research examines four different modern applications with each other for the first time (Twitter, Facebook, Email and Text message), it was decided to start with a small subset of features.

## 3.12 Classification Approaches

In the matching phase, the individual samples are compared with the reference template taken during the setup phase (i.e. the feature vector that results from the feature extraction process, which is clarified in section 4.4). Consequently, a match score is given to indicate the degree of their similarity, which decides the acceptance of the user's verification claim based upon the authentication decision.

As mentioned earlier, three different classification algorithms were examined to find the optimum algorithm for verifying message authorship: SVM, GB and RF. Each classifier was tested using a different set of features. They were selected due to most previous studies focusing on the classifier Support Vector Machine (SVM), and as it has achieved high performance in most modalities of authorship verification, as described in the literature review chapter, it has been employed in this research to investigate this classifier. While Gradient Boosting, and Random Forest classifiers were employed because they involve an accurate and effective procedure that allows the optimisation of an arbitrary loss function can be distinguished (Louppe et al, 2012; Singh et al, 2017); in addition, they were employed as they are modern classifiers used in data science, and to the best of the researcher's knowledge, no other study to date has used these two classifiers

for modern messaging platforms together. Chapter Five is next, and it provides more details about the classification results.

## 3.13 Conclusion

This chapter has presented the methodology that has enabled the investigation into authorship verification on the various platforms to facilitate a direct comparison of users' performance across different messaging platforms. It also discussed the opportunity to explore feature vector composition and the nature of classifiers to enable optimised performance. The methodology has included carefully considering the issue of privacy with respect to data being collected from these messaging platforms, as this would have been a major barrier to the successful completion of this study. The methodology included developing a privacy preserving mechanism to ensure no user data was captured during the process. One of the novelty of this section is the development and implementation of the privacy preserving data collection and feature extraction system. In addition, it was designed to investigate and analyse the message length required to enable reliable author verification decisions. Moreover, this increases the need for appropriate and sufficient information to create the reference template and perform verification and, importantly, it does not permit a direct comparison across systems.

To briefly describe this chapter, historical text message samples were recruited from the four core corpora of modern messaging systems of 50 participants, with the conditions that they had to have a least two out of the four identified platforms and were of consenting age (18 years +). A total of 275 features that included 227 stylometric, and 48 social network specific features with emoticon features, were extracted; of which, 48 additional features popularly used on social media such as emotional icons have been created and generated. In addition, each of the

messaging platforms required a process to be developed to parse the relevant messaging data, while ensuring only relevant data was parsed. For example, for Email, it was important to ensure the user's Emails were parsed and not the replies to Emails that are often appended to Email replies. Hundreds of scripts were improved and generated in order to perform a variety of functions to implement the pre-processing for each platform. An automated feature extraction program was also developed using NetBeans. This was built on a secure basis involving reading the individual input files confidentially without accessing the plain text exported from Twitter, Email, Facebook, and Text messages.

It is also important to investigate authorship verification in a platform independent manner, and to compare the relative performance of author verification across multi-short messaging platforms, including assessing how well author verification performs on individual platforms. In addition, exploring feature vector composition, as well as the impact of classification on performance, is necessary. Therefore, the next chapter presents platform independent author verification.

# 4. Chapter Five: Platform Independent Authorship Verification

This chapter presents the experiments that have been conducted in order to determine and compare the relative performance of author verification across short messaging platforms. It consists of two types of investigations into population and user-based verification approaches based on how feature vectors are composed. In developed biometric systems more generally, looking at feature vectors from a population perspective versus an individual user perspective is often referred to as static and dynamic feature vector composition; furthermore, it is necessary to determine the impact these approaches have on performance. From a biometric perspective, this chapter can be defined as presenting descriptive statistics that allow the nature of the data to be explored in order to find out what commonalities in data exist between platforms.The remainder of the chapter will be split into looking at the experiments based on population and individual user-based characteristics. Furthermore, this chapter describes the 17,280 experiments that needed to be conducted, bearing in mind that the dataset that was used is amongst the largest ever collected across platforms.

## 4.1 Population-Based Approach

A population profiling approach uses a population feature set for all those users examined for the specific platform dataset. A set of experiments were conducted with different settings to investigate the effectiveness of selecting different sets of features for verifying a given user's sample. The top features were captured for the population after ranking them by the (RF) algorithm, which included selecting 10 to 275 linguistic features as the input vector to a classification algorithm. To select these feature sets, a statistical approach was employed using the Random Forest algorithm.

### 4.1.1 Experimental Methodology

In a population-based approach, the process used to extract features from each platform in order to prioritise them in terms of discriminative information, prior to being applied to a standard supervised training methodology, has been described previously in (see section 4.6 Selecting Discriminative Features). The need to prioritise the feature vector is due to experimenting with the impact of feature vector length upon performance. Therefore, the Random Forest algorithm was used to deal with this as a multi-class classification problem. Only the top ranked features were fitted into the classifier in order to class them, based on a two-class problem for verifying them.

Secondly, in terms of data modelling, three different classification algorithms were used to find the optimum algorithm for verifying given message authorship: SVM, RF and GB. Each classifier was tested with a different set of features and train/test split ratios. This made it possible to achieve the desired goal of discovering the most important characteristics between platforms for a population-based approach.

Thirdly, the under-sampling technique has been used because a common problem that often occurs in authorship verification cases, is the lack of text samples to be used for training, as only limited text samples seem to be available for some authors. In contrast, large numbers of text samples may be available to other authors. Furthermore, text samples should be of comparable length; for example, some authors only have 20 samples while some others have over 19,000 samples, so it is not appropriate to make a comparison if the dataset contains imbalanced training set, this means the number of samples from positive class (minority) of the training dataset is much smaller than the number of examples of the negative class (majority) of the testing dataset (Ali et al., 2015).

Therefore, in order to solve this problem, the under-sampling technique that has been used in many research studies involves randomly selecting a subset of instances from the majority class and combining them with the minority class to form balanced class distribution data for model building (Stamatatos, 2008; Mukherjee et al., 2013; Gehrke et al., 2009). According to Stańczyk (2016), undersampling in data analysis is a good technique that can be used to adjust the class distribution of a data corpus. Hence, it has been employed in this research and applies to the training dataset not the testing dataset where all remaining samples are used and in order to handle and solve the class imbalance problem and to ensure classes are equal among all user samples. Finally, in order to reach the desired goal to find the most fitting features, a number of experiments were conducted, as shown in Table 4-1.

For each platform, each classifier was tested using 36 different sets of configurations (six features tested for each six train/test ratio) in order to investigate the most appropriate configuration with the most appropriate features, as illustrated in detail in Chapter Four, each of which was repeated by a number of users in the dataset by using a one-vs-all approach. It also involved splitting the datasets into the ratios of train/test changes for the system's performance (e.g. 70/30, 60/40, 40/60, 50/50, 20/80, 10/90 train/test), along with three classifiers, and the increasing top features tested were used (i.e.Test-1 to Test-36). This allowed the most appropriate characteristics to be selected and tested.

Table 4-1: Total number of tests for all datasets

| Platform | #Users | #Classifiers | Configuration | Total experiments |
|---|---|---|---|---|
| **Twitter** | 41 | 3 | 36 | 4,428 |
| **Text Message** | 26 | 3 | 36 | 2,808 |
| **Facebook** | 46 | 3 | 36 | 4,968 |
| **Email** | 47 | 3 | 36 | 5,076 |
| | | | ***Total** | 17,280 |

In this research, performance has been measured based on EER, which is used to evaluate the performance of classification algorithms (Jain et al., 2007). The experiments are based on 275 features sets, derived from five different types of categorised stylometric feature sets, which are lexical features (character-based features and word-based features), syntactic features, structural features, short message features, and emotional features. The type of feature sets and how they were selected has been described in detail in Chapter Four.

### 4.1.2 Experimental Results

These investigations have explored relative performance across platforms. In addition, examining the changing characteristics, along with classifiers and settings, has led to finding out the relative performance across platforms in order to select the best features, as most of the prior research has not described how to select the best sets of configurations or assessed relative performance across platforms. A number of scripts were written in order to perform a variety of tasks to implement the experiments (see Appendix G).

Table 4-2: Population-based experiment (one vs. all Authorship Verification)

| Test ID | Train/Test ratio | Feature tested | Performance EER (%) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Twitter | | | SMS | | | FB | | | Email | | |
| | | | SVM | GB | RF | SVM | GB | RF | SVM | GB | RF | SVM | GB | RF |
| Test 1 | 70/30 | Top 10 | 24.88 | 23.24 | 24.06 | 14.78 | 9.81 | 10.53 | 27.68 | 28.31 | 28.55 | 22.43 | 16.84 | 16.18 |
| Test 2 | | Top 20 | 24 | 21.03 | 25.51 | 13.67 | 8.56 | 10.78 | 27.97 | 27.25 | 31.97 | 22.43 | 13.31 | 15.61 |
| Test 3 | | Top 30 | 24.07 | **20.16** | 23.8 | 15.58 | 8.35 | 11.36 | 26.56 | 26.5 | 31.11 | 24.37 | 14.44 | 16.77 |
| Test 4 | | Top 50 | 25.78 | 20.77 | 27.3 | 15.9 | 8.19 | 11.42 | 27.89 | 26.69 | 28.42 | 24.8 | 13.65 | 17.09 |
| Test 5 | | Top 100 | 26.34 | 20.38 | 27.3 | 17.65 | **7.97** | 12.58 | 29.37 | 25.18 | 32.28 | 27.55 | **13.11** | 19.81 |
| Test 6 | | All | 31.41 | 20.47 | 29.37 | 21.11 | 8.1 | 13.82 | 38.44 | **25** | 33.39 | 32.78 | 13.5 | 22.71 |
| Test 7 | 60/40 | Top 10 | 24.09 | 22.66 | 25.25 | 15.1 | 9.9 | 10.58 | 29.03 | 28.97 | 29.3 | 22.51 | 17.16 | 16.77 |
| Test 8 | | Top 20 | 23.66 | 20.73 | 24.57 | 14.31 | 8.78 | 11.41 | 29.13 | 27.47 | 32.44 | 24.53 | 15.52 | 16.15 |
| Test 9 | | Top 30 | 24.78 | **20.28** | 25.5 | 15.3 | 8.42 | 12.11 | 27.77 | 27.06 | 30.84 | 23.67 | 15.51 | 17.68 |
| Test 10 | | Top 50 | 25.19 | 20.48 | 27.61 | 16.19 | 8.22 | 12.66 | 29.44 | 27.24 | 29.31 | 27.43 | 15.34 | 20.89 |
| Test 11 | | Top 100 | 27.28 | 20.91 | 29.91 | 19.59 | **8.13** | 13.1 | 29.86 | 26.52 | 32.98 | 28.57 | **15.68** | 23.88 |
| Test 12 | | All | 31.33 | 20.54 | 29.02 | 22.4 | 8.18 | 14.96 | 40.76 | **26.29** | 34.39 | 34.76 | 13.49 | 25.37 |
| Test 13 | 40/60 | Top 10 | 25.15 | 23.67 | 26.02 | 15.55 | 9.81 | 11.24 | 32.06 | 31.38 | 31.66 | 27.11 | 19.94 | 19.41 |
| Test 14 | | Top 20 | 24.56 | 22.44 | 26.76 | 15.54 | 8.72 | 13.28 | 32.25 | 31.02 | 35.74 | 30.53 | 19.46 | 18.91 |
| Test 15 | | Top 30 | 27.77 | 22.07 | 27.07 | 16.45 | 8.49 | 13.76 | 33.88 | 29.94 | 34.53 | 29.26 | 18.19 | 21.45 |
| Test 16 | | Top 50 | 31.18 | 21.89 | 29.93 | 17.37 | 8.42 | 14.47 | 37.62 | 30.89 | 32.97 | 28.96 | 18.24 | 23.64 |
| Test 17 | | Top 100 | 32.39 | 21.99 | 30.22 | 20.38 | **8.33** | 15.54 | 39.06 | **29.64** | 37.05 | 28.57 | **17.74** | 24.27 |
| Test 18 | | All | 34.86 | **21.76** | 31.94 | 23.17 | 8.39 | 15.98 | 42.65 | 29.99 | 37.57 | 39.57 | 18.47 | 25.73 |
| Test 19 | 50/50 | Top 10 | 24.1 | 22.87 | 24.47 | 15.46 | 9.74 | 11 | 29.04 | 30.66 | 31.14 | 23.32 | 17.4 | 17.34 |
| Test 20 | | Top 20 | 24 | 21.71 | 25.86 | 14.8 | 8.55 | 12.09 | 30.47 | 29.81 | 34.51 | 25.06 | 16.68 | 17.37 |
| Test 21 | | Top 30 | 28.76 | 21.18 | 25.86 | 15.91 | 8.34 | 12.48 | 31.5 | 28.94 | 33.77 | 28.74 | 15.69 | 18.97 |
| Test 22 | | Top 50 | 29.36 | 21.29 | 29.08 | 17.2 | 8.15 | 13.62 | 31.9 | 28.36 | 32.32 | 29.75 | 15.48 | 18.91 |
| Test 23 | | Top 100 | 31.11 | **21.11** | 29.86 | 19.83 | **8.02** | 14.71 | 35.65 | 28.2 | 34.52 | 31.91 | **14.34** | 24.13 |
| Test 24 | | All | 33.82 | 21.19 | 31.08 | 22.76 | 8.16 | 15.65 | 42.74 | **28.03** | 35.82 | 36.6 | 14.62 | 25.53 |
| Test 25 | 20/80 | Top 10 | 35.98 | 28.16 | 29.8 | 22.99 | 11.62 | 12.47 | 38.96 | 35.57 | 38.66 | 35.08 | 25.91 | 24.57 |
| Test 26 | | Top 20 | 35.63 | **26.76** | 32.05 | 23.02 | 11.41 | 13.94 | 40.18 | 35.73 | 39.07 | 36.3 | **24.72** | 25.59 |
| Test 27 | | Top 30 | 36.8 | 27.24 | 32.81 | 23.65 | 11.7 | 16.92 | 38.22 | 34.61 | 39.61 | 36.43 | 25.56 | 28.53 |
| Test 28 | | Top 50 | 35.88 | 28.84 | 32.44 | 24.7 | 11.59 | 18.8 | 40.61 | 35.23 | 37.17 | 37.54 | 24.77 | 28.4 |
| Test 29 | | Top 100 | 35.37 | 28.58 | 36.1 | 26.75 | **11.1** | 19.32 | 42.19 | **34.16** | 40.57 | 38.8 | 26.2 | 29.63 |
| Test 30 | | All | 37.78 | 28.87 | 35.29 | 29.07 | 11.36 | 20.86 | 48.77 | 34.58 | 39.78 | 50.26 | 26.36 | 33.03 |
| Test 31 | 10/90 | Top 10 | 39.29 | 33.9 | 34.5 | 24.62 | 12.35 | 14.89 | 49.25 | 39.52 | 41.64 | 48.55 | **26.92** | 27.93 |
| Test 32 | | Top 20 | 40.66 | **32.82** | 34.8 | 26.08 | **11.78** | 16.21 | 49.28 | 39.6 | 42.47 | 45.57 | 29.14 | 30.96 |
| Test 33 | | Top 30 | 41.94 | 33.7 | 36 | 26.19 | 11.93 | 18.04 | 49.64 | 39.67 | 42.56 | 46.12 | 29.85 | 32.53 |
| Test 34 | | Top 50 | 40.98 | 33.7 | 37.8 | 26.74 | 14.56 | 20.36 | 50.45 | 39.51 | 41.32 | 49.64 | 30.7 | 32.79 |
| Test 35 | | Top 100 | 43.4 | 34.88 | 38.49 | 29.81 | 15 | 19.6 | 52.04 | **39.35** | 43.62 | 50.89 | 31 | 37.3 |
| Test 36 | | All | 46.96 | 34.18 | 38.57 | 32.03 | 13.41 | 25.45 | 54.23 | 39.76 | 43.08 | 52.81 | 30.44 | 37.22 |

After performing recursive testing and a series of experiments were conducted as shown in Table 4-2, it was noted that the best features found in this experiment were for the Train/Test ratios 70/30 for all platforms: Twitter 20.16%, Text message 7.97%, FB 25.00% and Email 13.11% respectively, and the GB classifier showed the smallest EER on all four messaging platforms. From the experiments described above, one of the findings shows that increasing the features set often leads to the EER increasing. Interestingly, the worst results out of the platforms were for Facebook and Twitter, as shown in Table 4-2.

Several experiments were conducted to evaluate the forensic research question proposed by examining the reliability of population recognition when dealing with multiple messaging systems for the population base. Then the results from all platforms were analysed in the next trial as follows:

- The impact of the number of features on classification performance was investigated (i.e. top 10 features; top 20 features to top 275 features).
- The effectiveness of ratios of train/test changes on the system's performance (i.e.70/30 train/test) was tested.

More importantly, it has been noted that Text message and Email showed the best performance, and the best approach can be less than 100 features. This indicates that Text message and Email are more likely to have features that are similar among the population. The second reason for only the top 100 features or less being sufficient features for Text message and Email is the nature of these platforms, since they are private platforms and are not for puplic use; therefore, it is easy to distinguish one author from another due to the similarity of repeated writing styles and to obtain better performance.This is because it includes private or individual platforms for an author who usually has a single writing and unique

style, it is also typically one to one individual use, and a few of features are enough to discrimiante between authors/suspects.

In contrast, Twitter is considered to be public, with a small capacity, and writing ability is limited. The results show only the top 30 features is significantly sufficient to achieve a better performance; however, Twitter is used for puplic purposes, and the author is often writing for puplic for many different people, which may make it difficult for the classifer pick up and achieve high performance. There is another reason why this platform showed poor performance, which is because copy and pasted text messages between users on this platform is significant (Farahbakhsh et al., 2016; Ottoni et al.,2014).Therefore, based on performance, it may be possible to conclude that the advantage of privacy on these platforms (Text message and Email) can play a role in increasing performance and so causing them to outperform the Twitter platform.

Interstingly, the performance results for Twitter are better than Facebook, although both of them are being used in the public basis. However, Twitter has smaller capacity than Facebook and contains limited contents, that is, authors on Twitter must attempt a concise style to ensure their words are understandable and abbreviated, unlike Facebook, where the message can be large and unfocused (Russell, 2013). The difference between the performance of Facebook and Email, with Email outperforming Facebook, is likely because of the privacy issue, and as Email texts often tend to be more responsible. There is another reason why Facebook platform showed poor performance, similar to Twitter, which is because copy and pasted text messages between users on Facebook platform is significant (Farahbakhsh et al., 2016; Ottoni et al.,2014). This is the first study that has attempted to investigate the relative performance on multi-modern platforms together (to the best of the author's knowledge).

From a biometric perspective, a key requirement is the identification of potential settings for classifiers, and keeping only the best and most appopriate features. It can be observed that the best performance on all platforms for the population were 7.97% and 13.11% for the Text message and Email platforms respectively; compared to 20.16% and 25% for Twitter and Facebook respectively. Therefore, subsets of stylometric features would be more reliable in determining authorship for both the Text message and Email platforms. It can be concluded that during the population experiments, these results seem to indicate that there is a relative performance difference between the four modern platforms, generally ranked as follows: the Text Message platform would be more reliable for determining authorship and may be closer in terms of relative performance with Email, then the Twitter platform can be ranked next and is closer in terms of relative performance with the previous two platforms (Text message and Email). Finally, the Facebook platform is relatively far behind the previous composition platforms (Text message, Email, Twitter), so there is a real subset of common features between Email and Text message platforms. The GB classifier performed better on all messaging platforms. The next section provides user level performances to deepen the level of understanding of the relative performance between users across platforms.

### 4.1.3 Users' Performance Level Across Platforms

The section consists of the sets of further investigations conducted to address the core issues in the first set of research questions, which are related to understanding user performance on the messaging systems, and recognition using different message samples from the population base. Therefore, the users' performance was compared and explored across platforms. Table 4-3 below shows the user performance compared across four platforms for the population,

including the highest user EER and the lowest user EER for each platform. The differences between the highest authors' EER is in bright orange, while in contrast, the lowest users' EER is in blue.

Table 4-3: Authors' EER across Platforms

| Users | Twitter | SMS | FB | Email |
|---|---|---|---|---|
| **1** | 18.6 | 5.57 | 20.86 | 8 |
| **2** | 0 | 5.79 | 21.87 | 12.4 |
| **3** | 22.7 | 0 | 18.05 | 20.39 |
| **4** | 35.34 | 13.3 | 6.25 | 27.5 |
| **5** | 31.9 | 6.25 | 25.45 | 15.09 |
| **6** | 32.7 | 12.2 | 27.74 | 25 |
| **7** | 26.5 | 11.5 | 27.61 | 14.71 |
| **8** | 19.7 | 12.5 | 29.9 | 13.3 |
| **9** | 25.9 | 16.3 | 30.6 | 16.69 |
| **10** | 23.8 | 13 | 34.05 | 6.45 |
| **11** | 17.42 | 4.82 | 21.18 | 4.54 |
| **12** | 21.8 | 6.41 | 18.74 | 9.52 |
| **13** | 31.7 | 3.57 | 28.75 | 22.7 |
| **14** | 24.8 | 13.2 | 37.79 | 8.39 |
| **15** | 18.9 | 0 | 29.11 | 19.97 |
| **16** | 17.15 | 5.41 | 32.79 | 7.73 |
| **17** | 30.06 | 14.2 | 9.61 | 17.22 |
| **18** | 0 | 7.32 | 18.69 | 13.88 |
| **19** | 27.6 | – | 24.92 | 4.05 |
| **20** | – | 4.56 | 27.61 | 8.39 |
| **21** | 12.06 | – | 16.29 | 8.333 |
| **22** | 38.14 | – | 26.13 | 0 |
| **23** | 18.8 | – | 20.32 | 10 |
| **24** | – | 14.1 | 12.5 | 4.54 |
| **25** | – | 8.18 | 26.13 | 18.75 |
| **26** | 25.5 | – | 21.82 | 12.5 |
| **27** | – | 0 | 33.86 | 8.33 |
| **28** | 36.44 | – | 23.61 | 10.26 |
| **29** | 5.26 | – | 19.37 | 17.94 |
| **30** | 0 | 3.91 | – | 16 |
| **31** | 13.24 | – | 20.7 | 13.88 |
| **32** | 8.41 | – | 27.34 | 4.94 |
| **33** | 17.15 | – | 23.61 | 27.77 |
| **34** | 23.5 | – | 5 | 22.5 |
| **35** | 18.8 | – | 23.98 | 3.16 |
| **36** | 33.43 | – | 36.8 | 15.9 |
| **37** | 30.6 | – | 36.66 | 11.11 |
| **38** | – | 4.25 | 12.23 | – |
| **39** | 8.33 | 13.6 | 6.25 | – |
| **40** | – | 7.21 | 40.66 | 20.86 |
| **41** | 31.6 | – | 28.7 | 12.4 |
| **42** | 12.5 | – | 34.62 | 10.35 |
| **43** | 14.21 | – | 18.33 | 17.06 |
| **44** | 12.5 | – | – | 12.5 |
| **45** | 7.14 | – | – | 8.33 |
| **46** | – | – | 40.83 | 17.15 |
| **47** | – | – | 24.94 | 15.98 |
| **48** | 24.18 | – | 47.21 | – |
| **49** | 7.14 | – | – | 13.5 |
| **50** | – | – | 30.24 | 5.9 |

| | |
|---|---|
| | Highest User EER |
| | Lowest User EER |

When examining Table 4-3 for the users who have four platforms, it can be noticed that the performance for those users on all four platforms (i.e. *Users 1, 5, 6, 7, 8, 9, 10, 11,12,13 and 16*) are, gradually from best to worst, as follows: Text message, Email, Twitter and finally Facebook. This may infer that a pattern among those users exists. This was further verified across platforms based on the optimal features, which shows that there is a degree of similarity across platforms, starting with Text messages and then Email, Twitter and Facebook. This indicates that there were some common features shared between corpora for some authors. For example, the total number of character features (F1) and the total number of word features (F210) were examined later.

In comparison, it has also been noticed that for some users (i.e. *User 2, User 3, User 18* and *User 4*), there are differences and a more confused pattern, because the performance varies from platform to platform and is different from the previous relative performance. For example, as with the previous observation, it has been noticed that the best platforms vary gradually, starting from the best to the worst are as follows: Text message, Email, Twitter, and Facebook. However, for other users such as *User 2* and *User 18*, the best performance obtained was for Twitter, then Text message, Email and finally Facebook. Hence, the Twitter platform outperformed the Text message platform, and was followed by Text message, then Email and Facebook, which is almost the same pattern as in the previous example, although the difference in those users is only for Twitter, which outperformed the other platforms. Therefore, it can be inferred that the Twitter platform is closer and has been verified more strongly for *Users 2* and *18*; although the results are messy and differ for some other users concerning relative performance and the order of the platforms. This indicates that the Twitter

platform could be rich in similarity of features, with Text messages next, then Email and, finally, Facebook.

On other hand, it can be noted that when *User 3* was investigated, the best relative performance for them was Text message with an EER of 0%, Facebook with an EER of 18.05%, then Email with an EER of 20.39% and, finally, Twitter with an EER of 22.7%. This indicates that the Facebook platform had better performance, although this differs from the other users. A similar issue is apparent for *User 4*, as Facebook outperformed the other platforms with an EER of 6.25%, followed by Text message with an EER of 13.29%, Email with an EER of 27.5% and, finally, Twitter with an EER of 35.24% . While for *User 14*, the best relative performance are Email with an EER of 8.39%, Text message with an EER of 13.24%, Twitter with an EER of 24.8% and, finally, Facebook with an EER of 37.79%. In addition, for *User 15*, the best performance order is Text message with an EER of 0%, Twitter with an EER of 18.9%, Email with an EER of 19.97% and, finally, Facebook with an EER of 29.11% and for Iser 17, for the order is Facebook with an EER of 9.61%, Text message with an EER of 14.24%, Email with an EER of 17.22% and, finally, Twitter with an EER of 30.06%. Therefore, it can be inferred that the Text message and Email platforms are often at the forefront of platforms for the majority of users, due to often being first, or sometimes second.

From this point, it can be concluded when looking at Table 4-3 on user level performance, that relative performance is confused and inconsistent for some of the users who have four platforms combined with each other, and even those who have two or three platforms. However, it can be supposed that the best performance for most of users who have four platforms are, respectively, from the best to the worst: Text message, Email, Twitter and, finally, Facebook. The

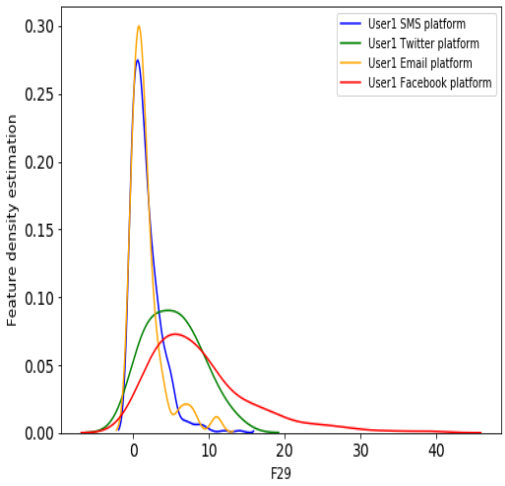
reason for the better performance and this order may be because there are several features that they both have common, as they work quite well independently; therefore, it can be suggested that lexical and syntactic features versus others could be the contributing factors regarding why they show stronger performance. That is, because from a linguistic perspective they have some quite unique categories of features, this means the performance characteristics are better when those features are used. Equally, these features may not work well with other platforms, for example Twitter and Facebook, leading to a fundamental impact, and this means the linguistic feature cannot be carried through four or more platforms.

There are limitations concerning the sample size when conducting research on big data with the goal of solving real life problems; for example, it's hard to predict what integration hurdles will be faced in practise with billions of online users if the recommendations from the research are taken forward. Despite as that, a high quality small sample can produce superior results and recommendations (Faraway et al., 2018), and scaling up is something that can be considered after the results have been analysed.

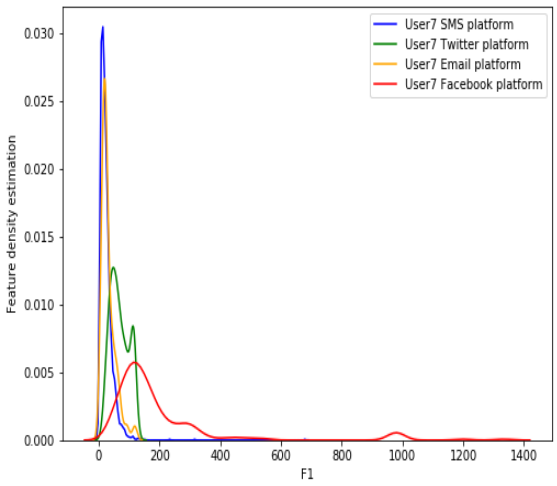
In order to see the similarities and differences between some linguistic features and some single user features across platforms, and in order to visualise the data and demonstrate the feature distribution and how features appear for author's across platforms, density estimation has been calculated. A density estimation tool can be used to view the data results of the authors, for example user feature distribution across platforms, and it can be used to visualisse the effect of discriminated features; the intensity of their use, and the differentiation between them. It has been used previously in the analysis of stylometric feature studies (Ding et al., 2017), and can create a smooth curve for a given set of data. In

addition, it can also be used to generate points that look like they have come from a certain dataset, such as features (Silverman, 2018). The distribution of the data shows the density function to a certain probability, and the density estimation allows the probability function to be estimated from the samples. The result is a function that represents the distribution of the data items in terms of their density in the data space. It is constructed on estimations based on noticeable data, from an unnoticeable underlying probability density function and can be utilised for threshold calculations. Ideally, it can be used for feature distribution since it has the ability to determine the differences in feature distribution for these top most importance features.

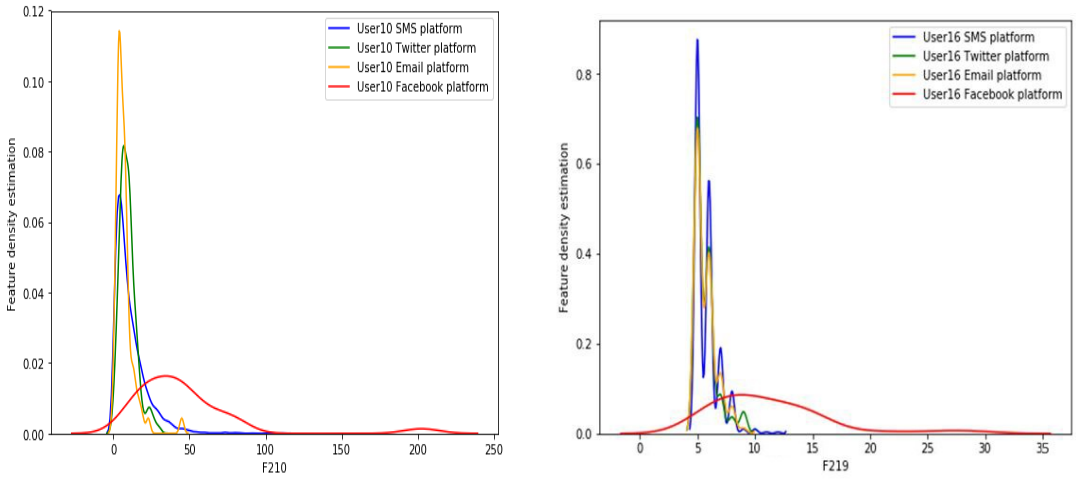
By performing density estimation calculations to explore the power of some features for a single user across platforms, it is possible to visualise the data and establish the degree to which input data is similar or dissimilar in feature distribution between platforms for the same user. Figure 4-1 shows the plots of density estimation to look for the similarities and differences between some categories distributed across platforms for some single users who have four platforms togather.




# number of alphabet a-z




# number of characters




# number of words

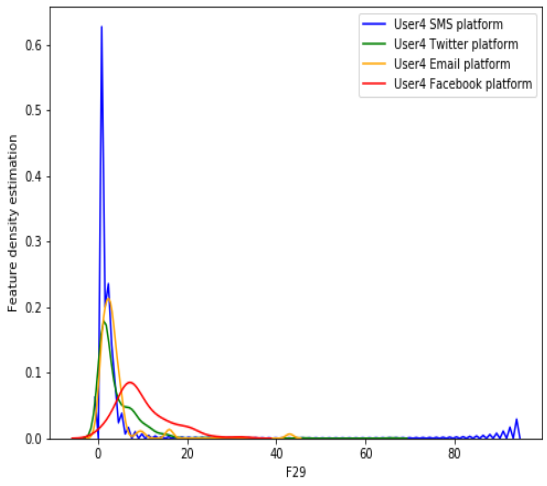
#number of words with 5 chars

Figure 4-1: Density estimation plots for similarity features for a single user across four platforms

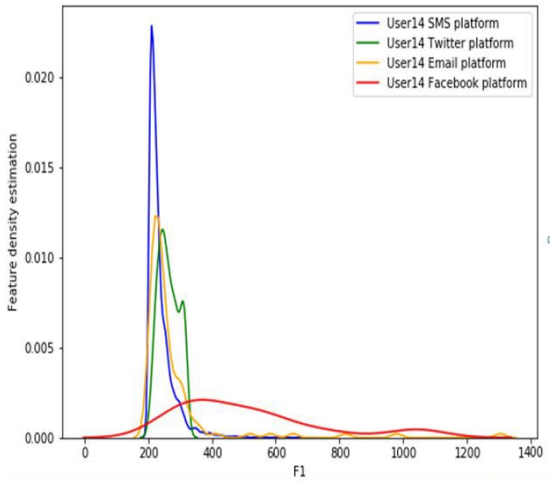
As mentioned earlier, in some cases, certain features performed well on a single platform, while they did not perform well on other platforms. One reason for the better performance on these two platforms could be because most of the uses of stylometric features, for example character and word-based features on Text message and Email are similar (Delany et al., 2012), and they contain language and conversations that are closer to a formal orientation, as well as often being addressed to a specific known person. For example, on average, User 1 may regularly use similar words and characters on the Text messge and Email platforms, resulting in EERs for Text messages of 5.57%, and 8.0% for Email. While on the Twitter and Facebook platforms, the EERs were 18.6% and 20.86% respectively, and they often differ. However, some features may be closer together in order of verification. Hence, further investigations have been conducted to understand relative user performance across platforms. Far from achieving a similar result in the verification compared to the previous two platforms are the Text message and Email platforms. This is unlike Twitter and Facebook where messages are often used and shared with people as they are social platforms; furthermore, the content is usually oriented towards the public

and may contain messages sent to unknown persons and copied from other authors. Therefore, most of the feature vectors of Email and Text message have similar characteristics, which is why the results for the performance of these are better than for the Twitter and Facebook platforms. In this sense, the pattern of the author can be determined more so using these two platforms based on the order of relative performance for those users mentioned earlier across platforms.

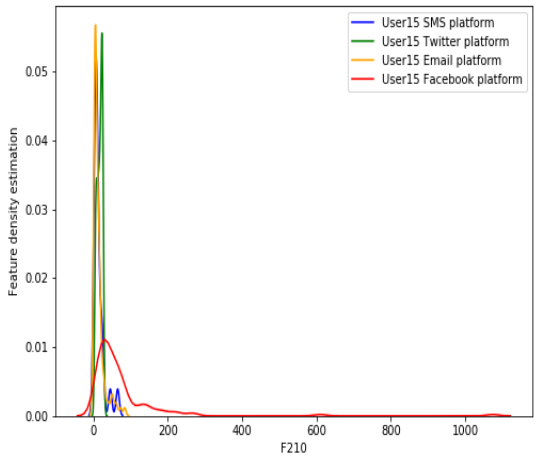
However, as mentioned earlier, there is confusion and the results are unclear when looking at some features for some users, especially the similarities and differences for some features for some single users. Having performed density estimation calculations to explore the differences for some features for a single user across platforms, Figure 4-2 shows the plots of density estimation to highlight the similarities and differences in distribution of some features across platforms for some individual users.

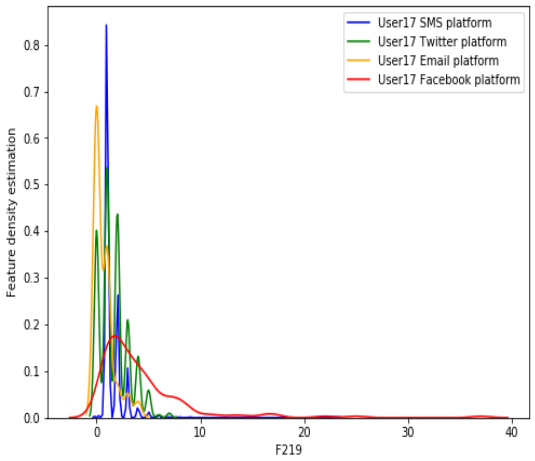



(a) # alphabet a-z




(b) # characters



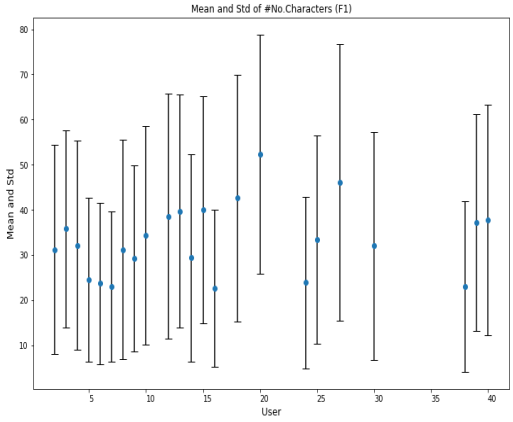




(c) # words (d) # words with 5 characters

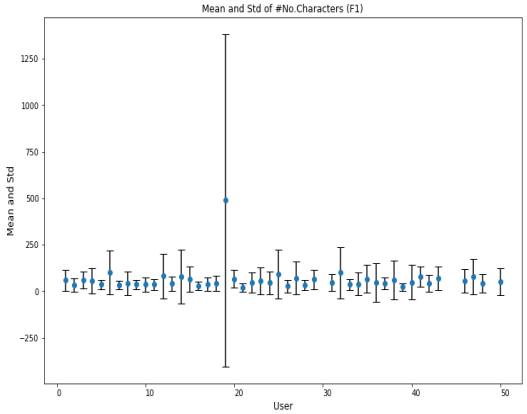
Figure 4-2: Density estimation plots for different features for a single user across platforms

Figure 4-2 shows the same features but for different users, and it can be noted that these features, which worked well with other users previously, as shown in Figure 4-1, do not distinguish these users across platforms; for example, *Users 4, 14, 15* and *17*. Further investigations have been conducted, including descriptive statistical analysis for some features, in order to look more closely at the similarites and differences between some features for some users. For example, the total number of character features (F1) and the total number of word features (F210) were examined to see their similarities and differences. By performing mean and standard deviation calculations, it is possible to establish the degree to which input data is similar or dissimilar between this group of users across platforms. Figure 4-3 and Figure 4-4 show the mean and standard deviation plot of these features, and the figures present each user's mean value, as well as the variance in character and word-based features, by calculating the standard deviation, to provide an estimate of the similarity and difference between the users' input vectors across platforms. Indeed, from the users' input data, two types of variance can be extracted: inter-class and intra-class variance. It is

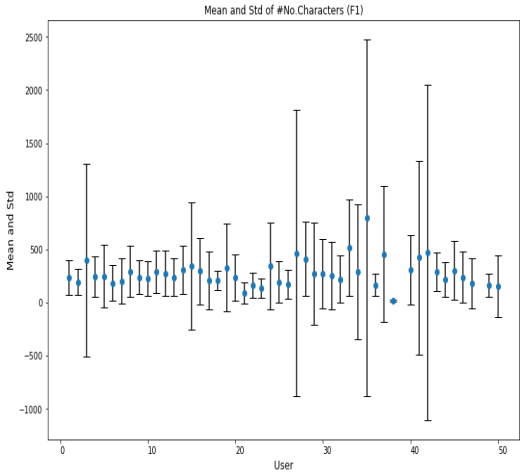
hypothesised that it would be easy to classify a user if the intra-class variance was ideally zero, so that every sample a user input would be identical, and the inter-class would be as large as possible, in order to widen the boundaries between features across platforms. The classification process across platforms seems to be much more complex, as the latency vectors observed from a single user is incorporated a fairly large spread of variance, which suggests the samples do not exist on a clearly definable discriminative region within and across platforms.



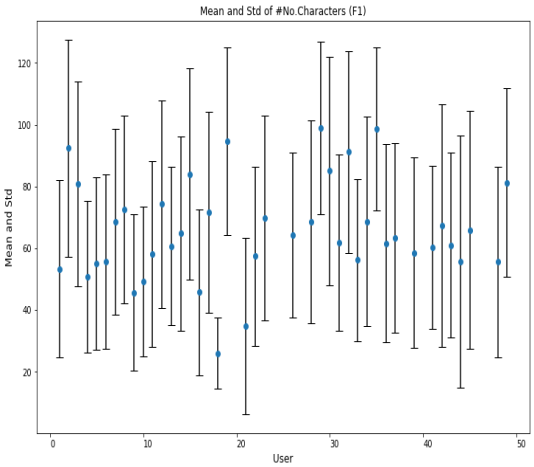

(a) # characters for Text message




(b) # characters for Facebook

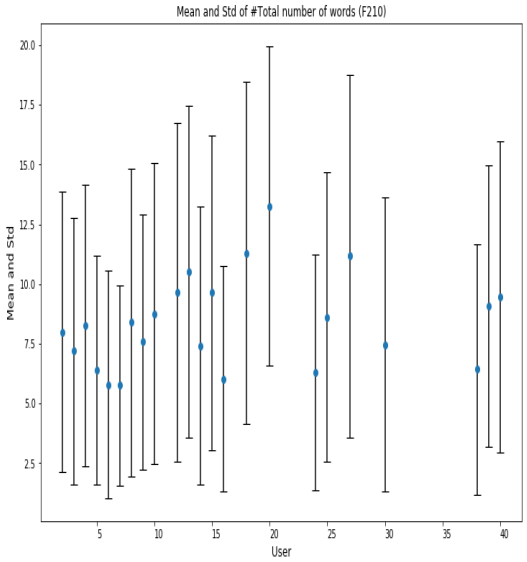



(c) # characters for Email

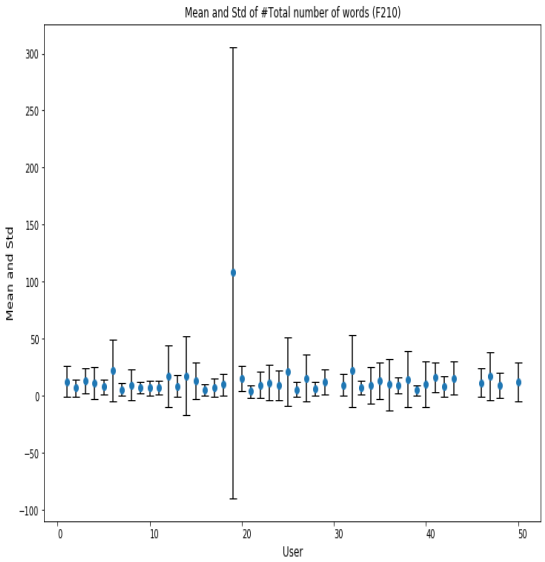



(d) # characters for Twitter

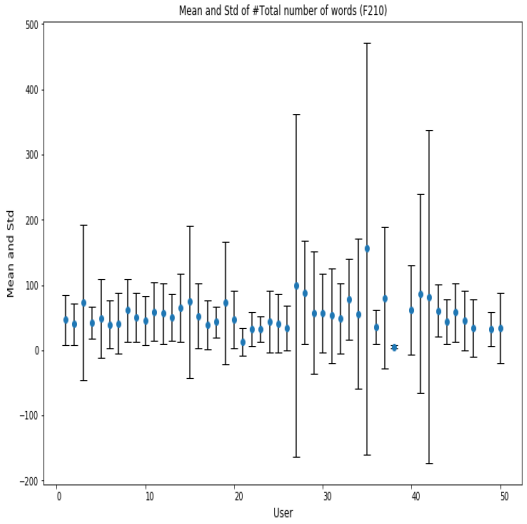
Figure 4-3: Mean & Standard deviation plot for character based features



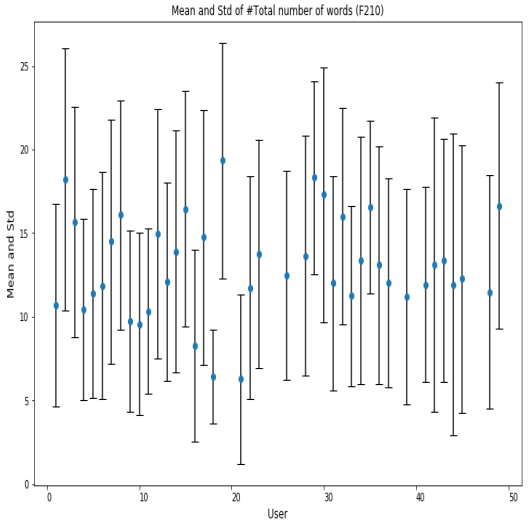

(a) # words for Text message




(b) # words for Facebook




(c) # words for Email



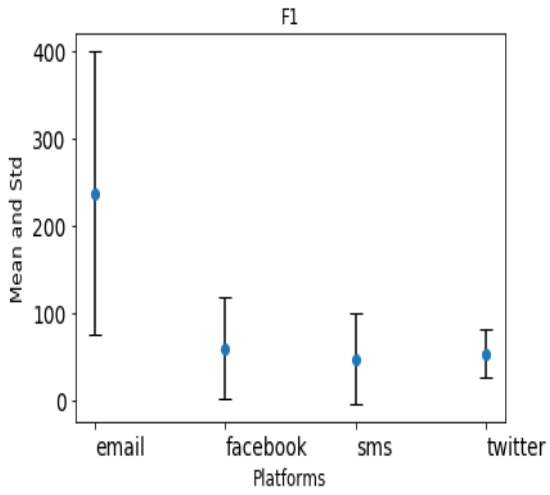

(d) # words for Twitter

Figure 4-4: Mean & Standard deviation plot for word based features

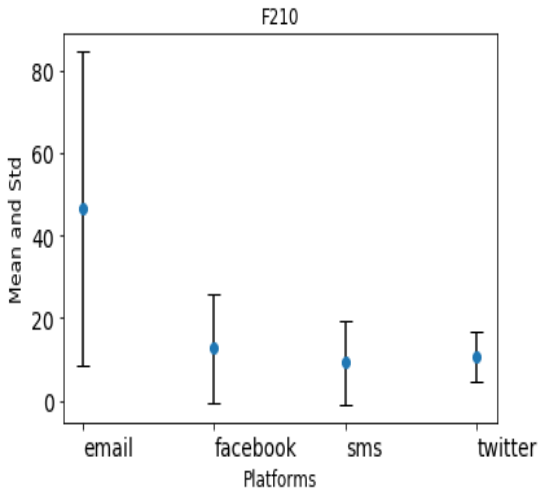
An initial analysis of the intra-class variance for these two robust features across platforms indicates that they are not ideal, as no users had a standard deviation close to zero for those two features across platforms; however, some users clearly have smaller intra-class variances than others. Furthermore, the majority of users have latency spreads that coincide with a number of others, demonstrating that they have low inter-class variance. This would definitely make

the classification more difficult as a user's input vectors are more likely to be similar, or within similar boundaries, to other users across platforms. Actually this data is quiet noisy, as there are some good examples of users where their intra is tidy and their inter is quiet large, that is, *User 19* and *User 20* on Twitter and Text message, but there are examples of features that are noisy and the intra of two users are similar, which is why the EER was higher.

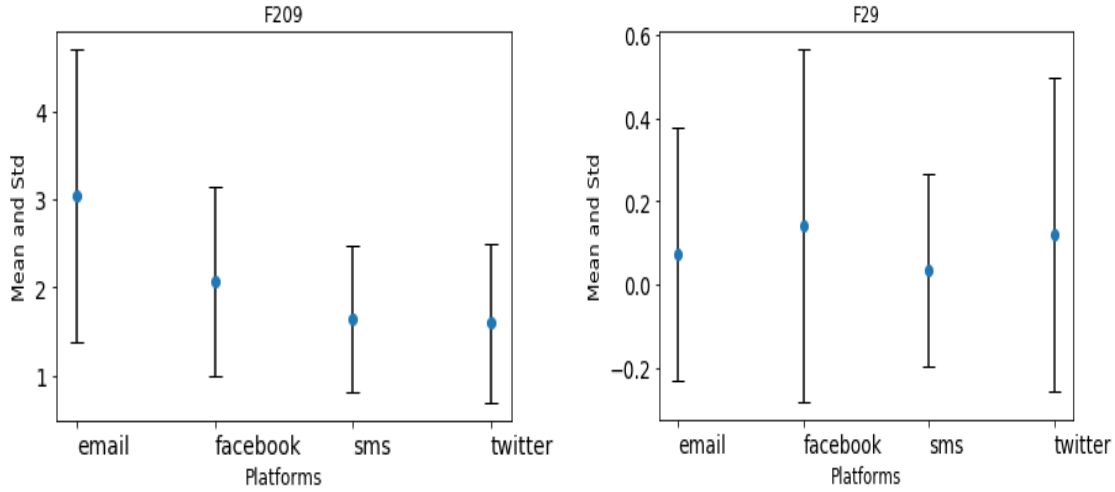
Further investigations have been conducted in order to explore the feature similarities of some users across platforms, and the inter-class variance was calculated to potentially illustrate the bigger picture. Figure 4-5 shows the mean and standard deviation plot for some features across platforms. The figure presents *User 1*'s mean value and also the variance in lexical character and word based features; number of sentences, and number of alphabet a-z features across platforms, from calculating the standard deviation. This provides an estimate of the similarity between features for the user input vectors across platforms.




(a) # characters




(b) # words




(c) #Sentence (d)# alphabet a-z

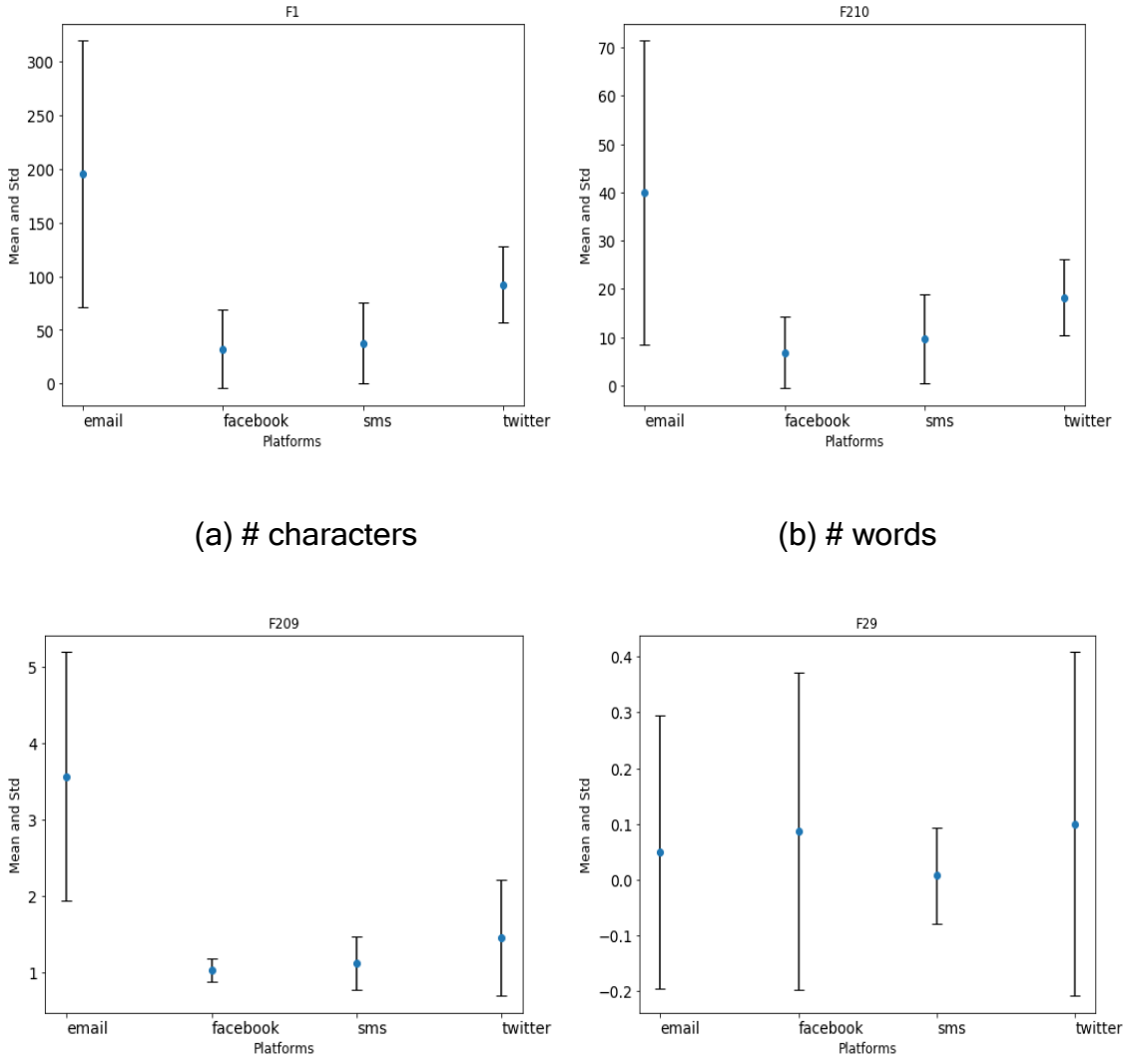
Figure 4-5: Mean & Standard deviation plot for some feature for User 1

Figure 4-5 shows the value from calculating the standard deviation for some features for *User 1* across platforms; the results are as follows: in a plot (a), feature # characters (F1): the Mean {236.6; 58.9; 46.9; 53.3}, and the Std {162.7; 58.1; 51.6; 28.5} for Email, Facebook, Text message and Twitter, respectively. While in a plot (b) # words (F210): Mean {46.5; 12.7; 9.1; 10.6}, and Std {38.0; 13.1; 10.1; 6.0} for Email, Facebook, Text message and Twitter, respectively. In a plot (c) #Sentence (F209): Mean {3.03; 2.07; 1.6; 1.5}, and Std { 1.6; 1.07; 0.83; 0.90} for Email, Facebook, Text message and Twitter, respectively. Finally, in a plot (d) # alphabet a-z( F29): Mean { 0.07; 0.14; 0.03; 0.11}, and Std {1.6; 1.07; 0.83; 0.90} for Email, Facebook, Text message and Twitter, respectively.

For the analysis based on the inter-class variance for *User 1,* some features differ (see Figure 4-5(a), (b) and (c)) across platforms, which indicates that the mean for plot (d) seems, generally, to be different from the rest of the platforms, indicating that the input vectors of this feature are not more likely to be similar between the mean features. For example, for the features on plots (a, b, c), their means are based on the most discriminating within platform as follows: Email,

Facebook, Twitter and Text message. While plot (d) is different, based on the most discriminating as follows: Facebook, Twitter, Email and Text message. It is possible that this feature distinguishes this user when using this feature compared to other features, as they differ from other platforms, as shown by plot (d) # alphabet a-z (F29): Mean {0.07; 0.14; 0.03; 0.11} for Email, Facebook, Text message and Twitter, respectively, and the nearest mean for the Email platform is the mean for the Text message platform, unlike other plots. In contrast, for other features within some of the plots, it is clear that in a number of sentences, the user's writing style is different across platforms (plot c) (structure feature). For example, Std of {3.03; 2.07; 1.6; 1.5} for Email, Facebook, Text message and Twitter, respectively. This perhaps seems normal because of the nature of Facebook and Email as they look similar concerning length of words for that user, unlike the Twitter and Text message platform, as can be seen from the plot. This also applies to the feature of # words, as the mean for that user is {46.5; 12.7; 9.1; 10.6} for Email, Facebook, Text message and Twitter, respectively; also notice that Twitter seems to contain more words than the Text message platform.It can be noticed that for User1, for example, the relative platforms for some features such as # alphabet a-z, F29 based on the mean value {0.03; 0.07; 0.11; 0.14} show gradual similarities in their performance, as follows: Text message, Email, Twitter and, finally, Facebook, respectively, for the means from the best to the worst performance, as shown in Table 4-3. This may suggest that the pattern for this user differs from the other confused patterns and could distinguish a particular group of user's performance from others, as shown in Table 4-3; in addition, there might a degree of similarity for this feature across the platforms. Therefore, another case study involving a different group of users is necessary to look at the differences between some features and patterns that cause the classifier to differ in its performance for those users that also have four

platforms. Figure 4-6 shows the mean and standard deviation plot for some features. It shows *User 2*s mean value and also the variance of lexical character and word based features, the number of sentences, and the number of alphabet a-z features across platforms, which have been reached by calculating the standard deviation to provide estimates of the similarities between features for the user input vectors across platforms.




(a) # characters (b) # words

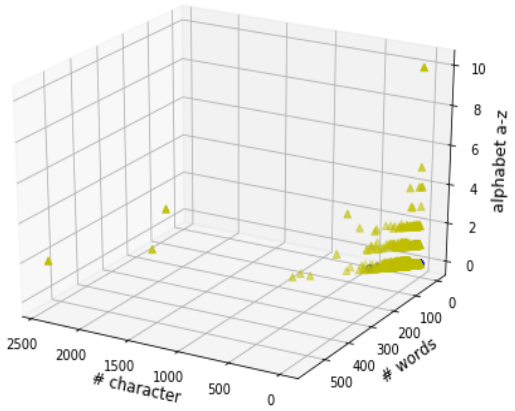
(c) #Sentence (d)# alphabet a-z

Figure 4-6: Mean & Standard deviation plot for some features for User 2 across 4 platforms

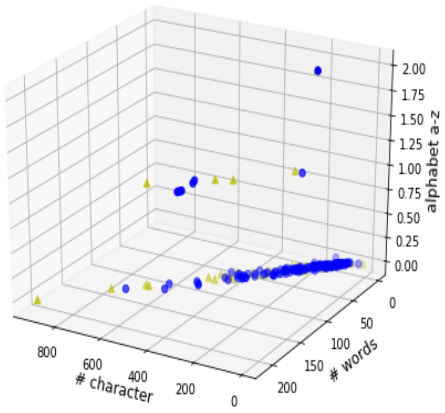
The value from calculating the standard deviation for some features for *User 2* across platforms is as follows: on plot (a), feature # characters (F1): Mean {195.2;

32.5; 37.8; 92.4}, and Std {0.3; 0.4; 0.2; 0.3} for Email, Facebook, Text message and Twitter, respectively. While for a plot (b) # words (F210): Mean {39.8; 6.8; 9.6; 18.2}, and Std {123.7; 36.4; 37.4; 35.2} for Email, Facebook, Text message and Twitter, respectively. While for plot (c) #Sentence (F209): Mean {3.5; 1.0; 1.1; 1.4 }, and Std { 31.5; 7.3; 9.2; 7.8} for Email, Facebook, Text message and Twitter, respectively, and for plot (d) # alphabet a-z ( F29): Mean { 0.04; 0.08; 0.007; 0.10}, and Std { 0.3; 0.4; 0.2; 0.3} for Email, Facebook, Text message and Twitter, respectively. An analysis of the inter-class variance for *User 2* for some features (see Figure 4-6(a), (b) and (c)) across platforms shows that the mean for the platform is generally different from the rest of the platforms, indicating that input vectors are not more likely to be similar between platforms. For example, the most discriminating platform using these features based on the mean value for features is the Email, Twitter, Text message and Facebook platforms, respectively, except for plot (d), as the value of the means are: Twitter, Facebook, Email and Text message, which is why the performance of Twitter was best for these kinds of users, as shown in Table 4-3.

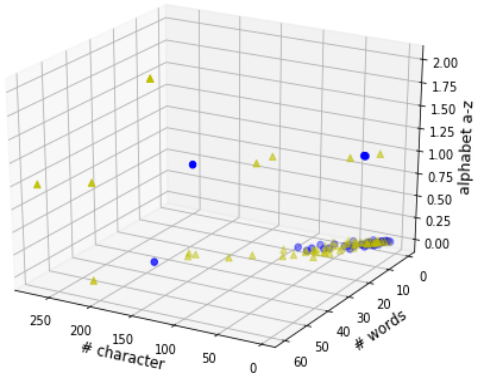
However, looking at the features to discover the similarities and differences in a single dimension across platforms does not convey the uniqueness and the nature of features that can be obtained through combining the features of users across platforms, making it necessary to move the discrimination into a multi-dimensional space. From a descriptive perspective, it is only possible to present three dimensions, and it is possible with two users having three features. Figure 4-7 shows the plot of two users and the number of characters, words and alphabet a-z data for each input vector. It can be seen as the differences are coded in colour between the two users (bright orange for *User 1* and blue for *User 2*).



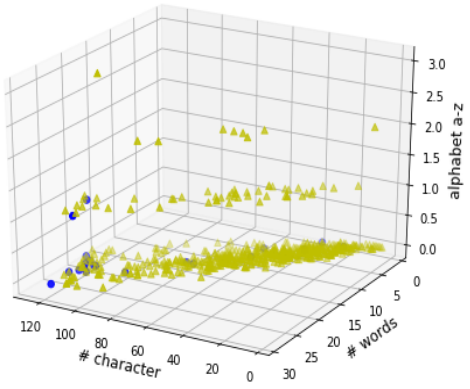

(a) *Users 1 and 2* in 3 features in Text message




(b) *User 1 and 2* in 3 features in Email




(c) *User 1 and 2* in 3 features in Facebook




(d) *User 1 and 2* in 3 features in Twitter

Figure 4-7: 3D plot of character and word- based features

Figure 4-7 shows a fairly complex problem with respect to deriving efficient decision-making boundaries. Assessing the discriminating features of users in a multi-dimensional space is not an easy task. It can be concluded that there might be some level of discriminative ability for Email, Facebook and somehow the Twitter platform. It might be possible to distinguish User 1 (bright orange colour) from User 2 (blue colour) by using these three features combined on plot b, then c and d. For example, it is possible for User 1 and User 2 (see Figure 5.5 (b), (c) and (d)) to be discriminated from each other according to these features as the

graph clearly shows that the area plots of data do not coincide with each other. While plot (a) clearly shows that it is difficult to differentiate between them according to these features. Therefore, the results are unclear when looking at certain features for some users who have four platforms, including the similarities and differences for all groups. In order to explore and understand the linguistic featues of each single platform more deeply, a series of investigations is presented in the next section. It is also suggested that there might be some level of discriminative ability for the linguistic feature category (d)# alphabet a-z (lexical- character feature) when used across combined electronic messaging systems for some users, and additional discriminative information could be provided to contribute towards boosting performance.

The most important and discriminating stylometric feature of each platform based on the highest EER and the worst EER for users have been established, and the details on their performances have been determined and provided. Investigating these discriminating features is required to understand the nature of the stylometric features of each platform, and in order to understand the relative performance for the reliable verification of Text message, Email, Facebook and Twitter users. In addition, the discrimination of the best of these stylometric features and their composition, which has affected the performance of each platform, is explained in detail in the following sections. Hence, a series of investigations were conducted, and the analysis of the feature vector composition between authors is presented in the next section.

### 4.1.4 Feature Vector Composition

There is no definitive science with a set problem that can lead to an absolute set of features or type of classifier, but it is a process subject to trial and iterations (Jain et al., 2000). Although many studies have attempted different stylometric

feature techniques using different corpora and different numbers of samples and authors, as shown in the Literature Review Chapter three, identifying the best and most appropriate features to facilitate author verification remains unclear. As demonstrated earlier, the feature sets that were extracted from the messaging systems are composed of 275 features. It is worth noting that the effectiveness of population features towards the classification cannot vary, and static features have a more significant impact on all authors' features, which were treated together. Therefore, a population-based feature approach was devised that can select the top 30 discriminated features, and to be more meaningful and use more specific features, the top thirty most important features were employed by the classifier across platforms. More importantly, identifying the most discriminatory features of these platforms based on uniqueness for population, may help the investigators of suspects to take into consideration the uniqueness of platforms, and how they differ from other corpus, to conduct reliable verification, as long as there are authors and a population that uses it.

It is important to define the most discriminative features for populations across each platform type. Therefore, this study has applied a large historical dataset containing 50 participants with strict procedures (at least 20 text messages must be obtainable from the user in one corpus). The most discriminative features for each platform across the population are presented in Appendix G, and the following sections illustrate the thirty most discriminative features across the population, including user level performance for each platform.

The analysis for each platform has been designed as follows:

- Analysis of the top six categories of features on each platform to better understand this feature category.

- Analysis of the top 30 features (a full listing of all 275 features can be found in Appendix G).

#### 4.1.4.1 Twitter platform

It may be beneficial to define the most discriminative population-based features for Twitter users. In order to find the most discriminating features for population on the Twitter platform, 4,428 tests were conducted with 13,616 samples of the historical data from 41 authors, as shown in Table 4-1. By using the GB classifier, it was possible to correctly classify the most discriminating features from the best performance achieved: an EER of 20.16%.

- **Top Most Important Features**

  In the Twitter experiments, the most effective features have been determined using the top 30 features for performance, yielding an EER of 20.16%, as shown in Table 4-2. The reason for determining the most significant features is to remove any redundant or irrelevant features, so that the classifiers would not be affected by noise, and the classification of new instances would be more accurate. It is important to determine the specifically required features to understand the performance of messaging systems for a population base. For example, if the results increased above the required features, without removing excess features, this would not improve the quality of the performance. While if the features were reduced, the quality would be insufficient to perform its task. Therefore, the results show the top 30 features for the Twitter platform, and the best performance and optimal threshold can be obtained using these features. GB was the best and most appropriate classifier, with data split 70/30 for training/

testing. A summary of the top features with EER is shown below in Figure 4-8.

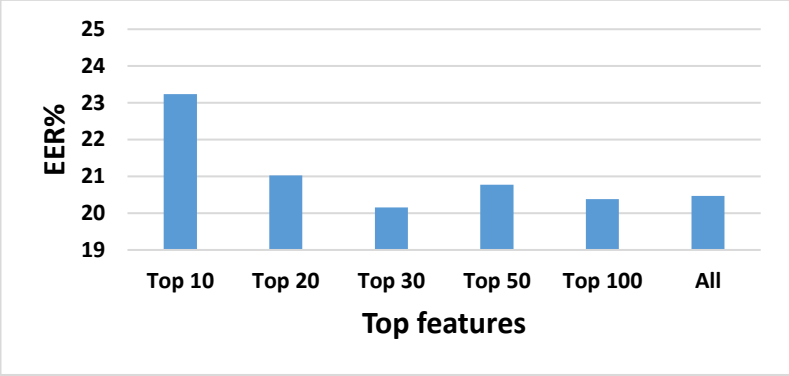



Figure 4-8: Top features with EER for the Twitter platform

Table 4-4 shows the top 30 features on the Twitter platform. In order to answer the research question: What commonalities and differences might exist within the feature set for population?, the feature vector and how it impacts performance was investigated; in addition, an analysis of the similarities and differences in feature vectors between authors on the Twitter platform is provided. The procedure of analysis will involve the first Top 30 features. It should be noted that in order to further understand what these features are top for on each single platform, the next chapter presents a qualitative comparison, and shows how these are important to each other, after conducting a comparison of each case.

Table 4-4: Top Discriminative Features in a population for Twitter

| Twitter Top discriminative Features in Twitter | | | |
|---|---|---|---|
| **No** | **Top 30 Features** | **Features category** | **Feature name** |
| 1 | 31 | (Lexical) | Number of special character. ("@"). |
| 2 | 231 | (Short Message features) | Frequency of missing an uppercase letter when starting a sentence. |
| 3 | 55 | (Syntactic) | Number of punctuation. (":"). |
| 4 | 3 | (Lexical) | Number of uppercase characters. |
| 5 | 1 | (Lexical) | Number of characters. |
| 6 | 2 | (Lexical) | Number of alphabets. |
| 7 | 52 | (Syntactic) | Number of punctuation. (".."). |
| 8 | 54 | (Syntactic | Number of punctuation. ("!"). |
| 9 | 213 | (Lexical) | Average sentence length in terms of character. |
| 10 | 39 | (Lexical) | Number of special character. ("_"). |
| 11 | 32 | (Lexial) | Number of special character. ("#"). |
| 12 | 48 | (Lexical) | Number of special character. ("/") . |
| 13 | 210 | (Lexical) | Total number of words. |
| 14 | 214 | (Lexical) | Average sentence length in terms of word. |
| 15 | 27 | (Lexical) | Number of alphabet a-z. ("x"). |
| 16 | 212 | (Lexical) | Average word length. |
| 17 | 227 | (Lexical) | Number of words with more than 12 chars. |
| 18 | 219 | (Lexical) | Number of words with 5 chars. |
| 19 | 209 | (Structure) | Total number of sentences. |
| 20 | 23 | (Lexical) | Number of alphabet a-z. ("t"). |
| 21 | 21 | (Lexical) | Number of alphabet a-z. ("r"). |
| 22 | 232 | (Short Messages feature) | Frequency of missing a period or other punctuation to end a sentence. |
| 23 | 8 | (Lexical) | Number of alphabet a-z. ("e"). |
| 24 | 22 | (Lexical) | Number of alphabet a-z. ("s"). |
| 25 | 224 | (Lexical) | Number of words with 10 chars. |
| 26 | 233 | (Short Messages feature) | Frequency of missing the word “I” or “We” when starting a sentence. |
| 27 | 51 | (Syntactic) | Number of punctuation. (","). |
| 28 | 211 | (Lexical) | Total number of short words (less than four characters). |
| 29 | 4 | (Lexical) | Number of alphabet a-z. ("a"). |
| 30 | 19 | (Lexical) | Number of alphabet a-z. ("p"). |

| (Lexical) | (Syntactic ) | ( Short Messages) |
|---|---|---|
| Top Repeated | Second Repeated | Third Repeated |

As demonstrated in Table 4-4, the most repeated features category used when the top thirty features were captured are as follows: Lexical features were repeated over 21 times, followed by syntactic features four times, and finally, short message features three times, while structure appeared once. For the

lexical features, it can be noted that the number of alphabet a-z appeared seven times, and the number of special characters appeared four times – a special character feature also appeared at the top of the list. As it can be seen, the lexical features were repeated and covered almost more than half of the dataset when the top 30 features were examined, and the most distinguished features of the lexical category is the number of special characters, since they are repeated four times, and the feature number of alphabet a-z, since this appeared seven times. In second place came syntactic features, due to being repeated three times, especially the feature 'number of punctuation marks.' Therefore, the number of special characters, number of alphabet a-z for lexical type and number of punctuation marks for syntactic features seem to be the most distinctive on the Twitter platform when the top thirty features are established, and this helped improve performance.

In general, when the top 30 features were investigated for the Twitter platform, lexical features were the most discriminated, as shown in Table 4-4. This comprehensive review shows the top features when the top 30 features of the dataset are captured, and further investigation will be presented in the next sections to show what commonalities and differences exist between populations for these features, including for the top most categories.

Fundamentally, this section seeks to present an understanding of the impact of population feature techniques and the value of feature space on the performance of each platform (i.e., the commonalities and differences that exist within the feature sets and across platforms). It also seeks to explore the classification performance of the population on multi-platforms.

To investigate the effect of the top most discriminating features between the population on the Twitter platform, an experimental analysis of author features is

provided in order to present a comprehensive picture and better understand the nature of these top most discriminatory features as an input vector between users. This has led to the most significant discriminative features for authors on the Twitter platform being ranked in order, and to answer the part of research question on what commonalities and differences exist. User level performance and feature distribution are provided in the next subsections.

Indeed, the analysis of the most discriminate features for population includes investigating the author level; fundamentally, the features need to focus on two particular biometric characteristics: their capability to be universal and their uniqueness. In order to select an effective and universal subset of features for individual platforms that can aid author verification, most stylometric features and social network specific features were employed in this study.

From a biometric perspective, with respect to being universal (e.g. lexical feature appears over 21 times more than half the features in the top 30 for the Twitter platform), the type of lexical feature plays a significant role in improving the performance of the Twitter platform, as shown in Table 4-4 above. Secondly, with respect to the unique features (e.g. determining which robust features categorise this platform) that can be used for discriminating authors. The aim is to present some useful insights into the identity of the author features in this platform, and establishing the extent to which the input data is similar or dissimilar between authors is significant to investigate the ability to determine potentially positive features for forming unique patterns to distinguish individual authors. There are two types of groups of authors concerning performance, as can be seen in Table 4-5. The first type of authors showed good performance, for example, the best case of individual performance achieved an EER of 0%. In contrast, the second group of authors showed poor performance, for example, the worst case of

individual performance achieved an EER of 38.14%. Thus, the top 30 features are powerful for discriminating between populations. In general, poor author performance due to author input vector features differ in writing style, making it difficult for a classifier, since Twitter is mainly used for public viewing (Ashcroft et al., 2015). In order for the top features to be discussed for each platform individually, the *highest author's EER* and the *lowest user's EER* have been selected as this gives an approximate description of the most common features among users. Table 4-5 shows how the top 30 features are distributed across the population, including the highest user EER and the lowest user EER. The differences between the highest author's EER (bright orange colour) for *Author 22* can be seen, while in contrast, the lowest user's EER (blue colour) is for *Author 30*. Therefore, a series of investigations were conducted, and the analysis of feature vector distribution between authors is presented in the next section.

Table 4-5: Authors' EER for theTwitter platform

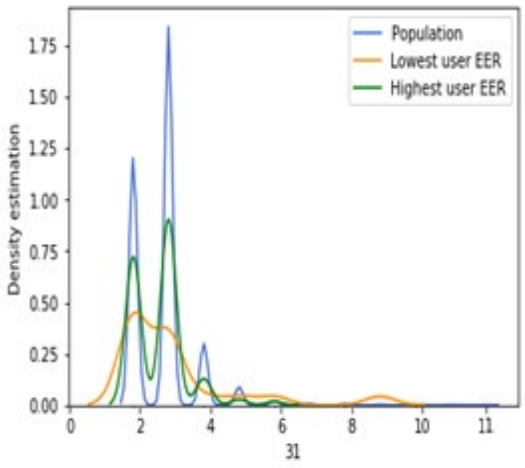
| Users | EER | Users | EER |
|---|---|---|---|
| **1** | 18.6 | **23** | 18.8 |
| **2** | 0 | **26** | 25.5 |
| **3** | 22.7 | **28** | 36.44 |
| **4** | 35.34 | **29** | 5.26 |
| **5** | 31.9 | **30** | 0 |
| **6** | 32.7 | **31** | 13.24 |
| **7** | 26.5 | **32** | 8.41 |
| **8** | 19.7 | **33** | 17.15 |
| **9** | 25.9 | **34** | 23.5 |
| **10** | 23.8 | **35** | 18.8 |
| **11** | 17.42 | **36** | 33.43 |
| **12** | 21.8 | **37** | 30.6 |
| **13** | 31.7 | **39** | 8.33 |
| **14** | 24.8 | **41** | 31.6 |
| **15** | 18.9 | **42** | 12.5 |
| **16** | 17.15 | **43** | 14.21 |
| **17** | 30.06 | **44** | 12.5 |
| **18** | 0 | **45** | 7.14 |
| **19** | 27.6 | **48** | 24.18 |
| **21** | 12.06 | **49** | 7.14 |
| **22** | 38.14 | | |

Highest User EER
Lowest User EER

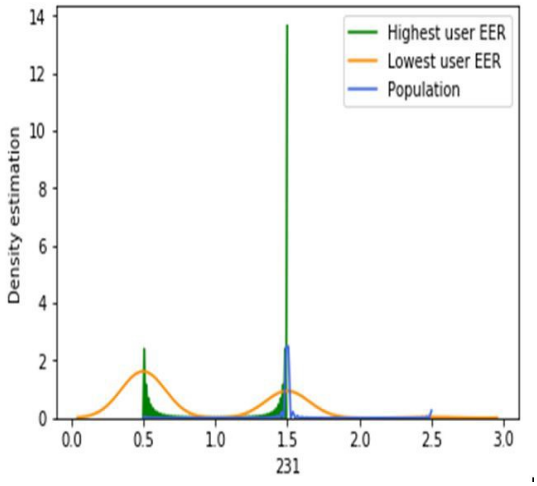
## ➢ Examining the Top Six Categories

Different stylometric features have been used that cover a wide range of writing styles on the Twitter platform, as described in the Research Methodologies Chapter, including features such as social icons (a full listing of social icon features can be found in Appendix B), which, have not been employed in most literary studies. It has been explored and expected that the feature of social icons would contribute to a social platform like Twitter, however, the results show that this feature did not appear among the top six categories of the Twitter platform. Therefore, as expected, the top six most discriminating features for the Twitter platform was lexical, as that covers most of the top six categories.

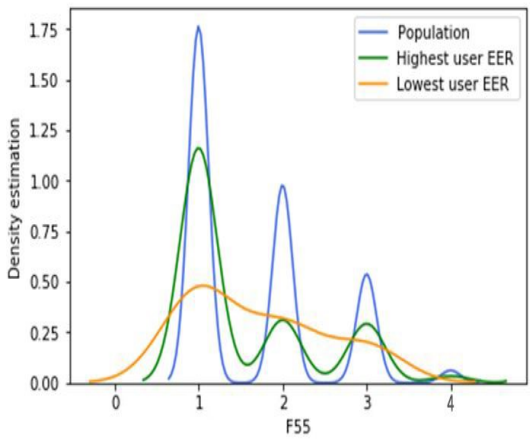
Figure 4-9 shows the plots of the top six categories. It can be seen that the differences are coded in colour between the highest user EER (green colour), population (blue colour) and lowest user EER (dark orange). From the previous table, Table 4-5 on user EER, the lowest user EER (Author 30) can be seen to be different from the lowest error rate (Author 22) and different from other populations.



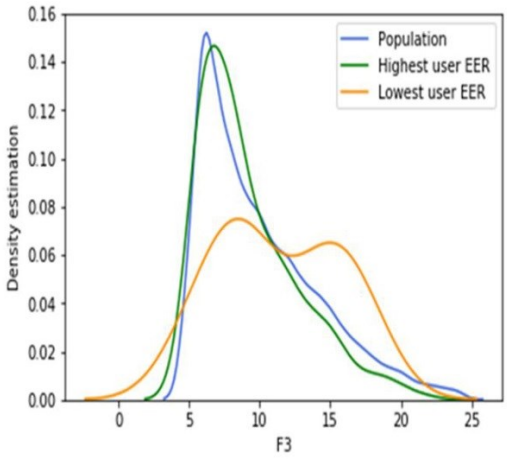

(a) #Special character. ("@"), (lexical).



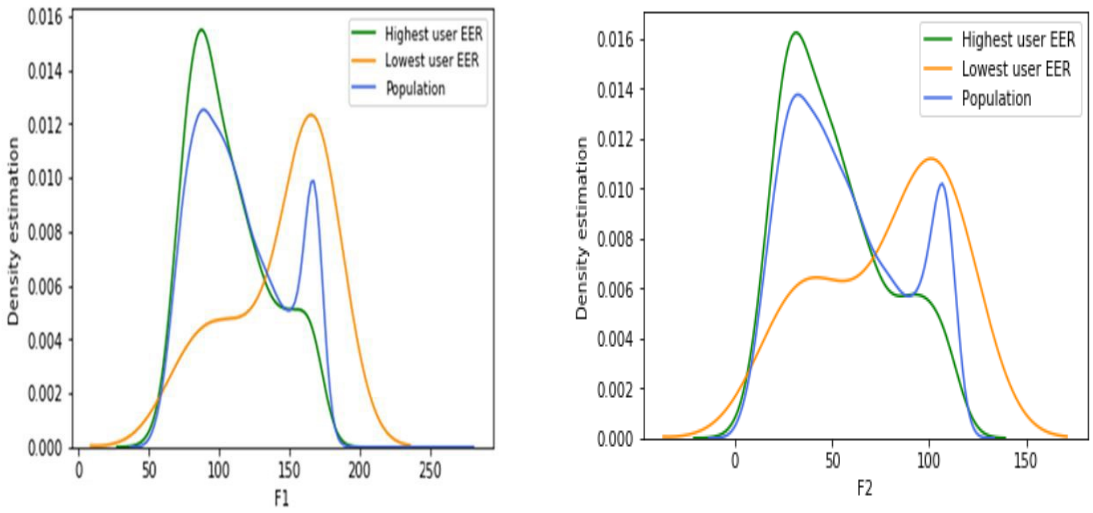

(b) #Missing an uppercase letter, (Specific feature)




(c) # Punctuation (":"), (Syntactic).




(d)# Uppercase, (lexical).




(e) # Characters, (Lexical). (f) # Alphabets, (Lexical).

Figure 4-9: Density estimation plot for top category features on Twitter

Figure 4-9 shows that the feature 'lexical' was top from amongst other categories for the population on the Twitter platform as shown in the previousTable 4-4. This feature was able to distinguish one author from another. It shows that lexical features covered most categories and that includes #special character F31, # Uppercase F3, # Characters F1, and #Alphabets, F2 , which for the lowest author error rate was (0) (*Author 30)*; the highest user error rate was (38.14) (*Author 22)* as shown in Table 4-5 for example for *Author 22* and the *population*. The total number of special characters, number of uppercase characters, number of characters, and number of alphabets categories on the Twitter platform played an important role in discriminating authors. The graph shows the similarities and differences in the input data between authors and the population, although most of the population use this feature; however, this category of lexical features has the ability to distinguish between authors/suspects. The *population* does not share similar feature vectors with *Author 30* with a #special character feature. While, as can be seen, the highest user ERR (*Author 22)* and *population* are almost similar, and this difference distinguishes *Author 30* from the *population*, which is due to their input vectors not being similar to the rest of the *population*.

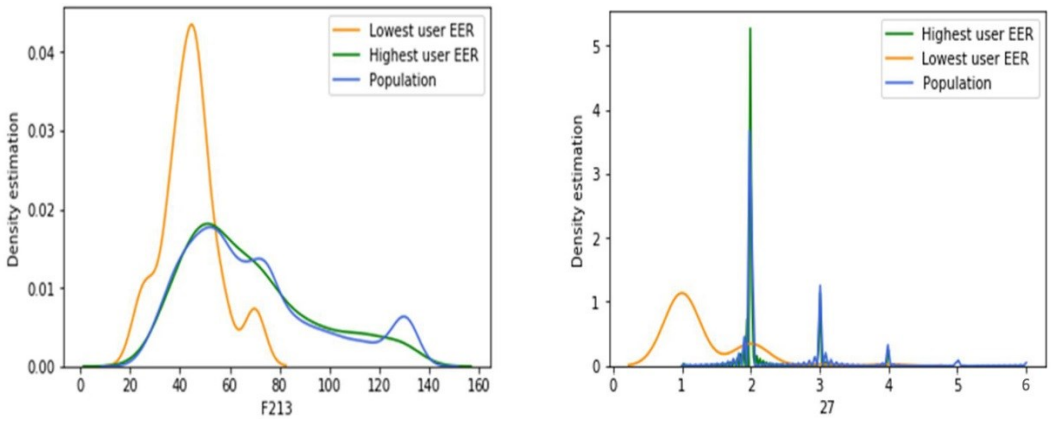
Linguistically, it can be noted that some Twitter users may use special characters specifically on these kinds of platforms because Twitter, Text message, and so on, have a limited number of words and do not allow more words. This indicates that some users tend to use shortcuts such as "$, @, #, %, ^ ,& ,* as they are a brief and express what the person wants to say in a way that is understandable to others. In addition, some of these symbols have a specific goal for their use, for example the sign "@" is used to mention someone on Twitter and it is functional on this platform; thus, it can be suggested that most people use it in their tweets to draw attention to the receiver or if their point needs to be retweeted (Almishari et al., 2014). On the other hand, it can be suggested that some users do not use this feature "@" because they are famous or have many followers - maybe hundreds of thousands of followers - so they do not need to mention anyone else since their followers will retweet it, and so it will reach a large number of people, even the intended person. This is possibly why some people use it, especially on the Twitter platform, and it has helped in distinguishing users in this experiment. As a result, these categories are robust categories of lexical features on the Twitter platform, since it has the ability to distinguish author/suspects, and it differs from other platforms, making it a unique feature of this platform that is useful for the investigation of suspects. To further investigate other features on the Twitter platform, and in order to examine other types of features such as syntactic, structure, and so on, it was repeated and appeared in the top thirty most important features, as shown in Table 4-4. Therefore, the next section provides a set of investigations of feature categories to deepen the level of understanding of differentiation and discrimination for the population.

### ➢ Examining the First Top Thirty Important Features

#### • Lexical features

This section will examine in detail the way in which these types of features have an impact on the performance of Twitter, and to establish the extent to which input data is similar or dissimilar between authors on the Twitter platform. As illustrated previously, the feature *number of special characters* has the ability to discriminate between authors and is active on this platform since it has discrimination information, which has been explained in detail and is the top most feature, as shown in Table 4-4. Some other categories of lexical, syntactic, structure and short messages features have also been investigated. As illustrated previously in Table 4-5, the lowest user error rate was (0) (*Author 30)*, while in contrast, the highest user error rate was (38.14) (*Author 22)*. Therefore, the investigation of other top lexical categroies included: average sentence length in terms of character, and number of alphabet a-z, F213 and 27, as shown in Table 4-4. They have been investigated because it was shown that they were the most important lexical features from amongst others on the Twitter platform.

To visualise the data, and demonstrate feature vector distribution between authors and establish the degree to which input data is similar or dissimilar between authors. Figure 4-10 shows the plots of a univariate kernel density estimation to determine the differences in feature distribution for the top most important distribution lexical features between authors, with density estimating showing the degree of discrimination for the *population*, and the lowest user EER (*Author 30)* and highest user EER (*Author 22)*.




(a) Average sentence length in terms of characters (b) # number of alphabet a-z

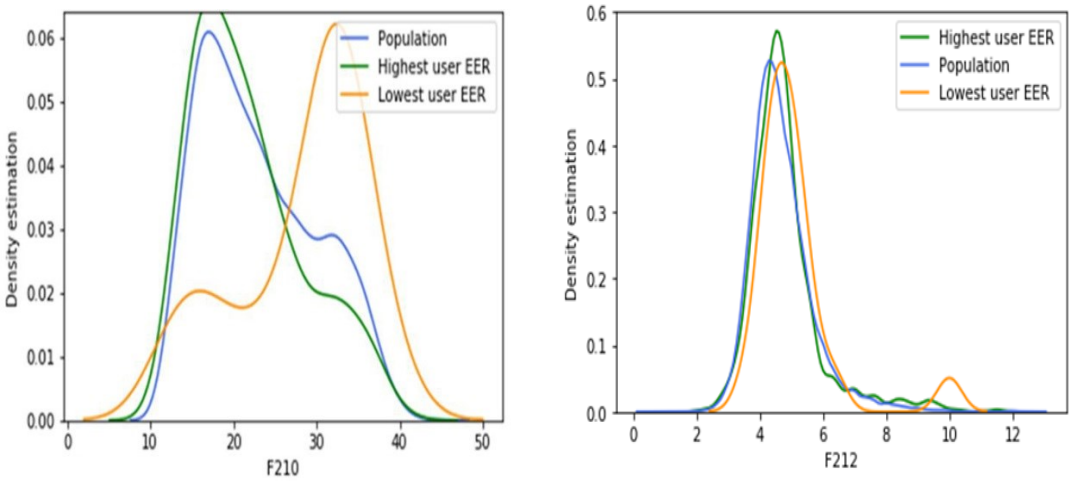
Figure 4-10: Density estimation plot for lexical categories on Twitter

In plot (a) *Average sentence length in terms of character F213*, *Author 30* uses this feature, and mostly between the range of 20 to 80 for average sentence length in terms of characters. The data on *Author 30* is concentrated on this feature, peaking between 20 and 40 for his/her average sentence length in terms of characters, which means that *Author 30* often uses this sentence length on the Twitter platform. It can also be noted that *Author 22* is difficult to discriminate from the *population* as they clearly share a similar boundary with the *population* who use this feature significantly - between 0 and 140. Therefore, this indicates this feature provides some level of discrimination, making it somewhat effective and active on this platform because it is possible for *Author 30* and *Author 22* to be discriminated from each other, as the graph clearly shows that the area plots of data do not coincide with each other, and Author 30 does not coincide with the *population*.

Overall, the features *Average sentence length in terms of character F213*, and feature *# number of alphabet a-z F27* can have some level of ability to discriminate *Author 30* from *Author 22,* and to some extent, the *population*.

This section illustrates that lexical character features can have some level of discriminative ability for some authors on the Twitter platform. Lexical character-based features can certainly distinguish authors from each other due to their input vectors not being similar or not being within the same entire boundaries as the *population*. Thus, lexical feature has become one of the most important types of stylometric features. However, plot (b) shows that it may be rather difficult to discriminate *Author 30* from the *population*, although they can be discriminated by a chance of 10%, because as the graph shows, the area plot of data coincides between *Author 22* and the *population*.

To examine the way in which authors use short or long words on the Twitter platform (this feature is explored more in Section 6.5), the distribution of word number and average word length for the *population* has been examined. The lowest author error rate was (0) (*Authors 30)*, and the highest author error rate was (38.14) (*Author 22)*. Figure 4-11 illustrates the number of words and their distribution.




a) Total # words (b) Average word length

Figure 4-11: Density estimation plot for total number of words and average word length features

Further analysis has been conducted, and an analysis of the *number of word* feature indicates that the majority of authors on Twitter used *#words* that were an average of 10-40 words long in their tweets, as can be seen in plot (a). However, as can be noted in plot (b), most of the *population* tend to use short words on the Twitter platform, centred between 2-8 characters. This is expected, as authors have to find a way of being concise and short in their tweet messages with a limited number of words. In contrast, it can be also noted that *Author 22* (the highest author error rate) uses more words, centred on around 35 words, than the *population*, which is significantly larger when compared to the *population*; therefore, this feature definitely distinguishes *Author 22* from others. Therefore, this feature can be robust and provides discriminative information, with a level of discriminative ability, on the Twitter platform, and it is among the top 30 discriminatory features, as shown in Table 4-4.

Further analysis of some of the most discriminating features from among the top 30 most important features has been conducted. Plot (b) describes the average word length distribution usage for the *population*, the lowest user EER (*Author 30)* and the highest user EER *(Author 22)*. It can be noticed that most of the *population* tend to use words in their tweets that are an average of five characters long; yet a difference is that *Author 30* centeres around 10 characters, which distinguishes them from the *population*. However, overall, this measure does not provide a robust feature because *Author 30* cannot be clearly discriminated from the *population*, and the data indicates that input vectors are more likely to be similar between authors, as the graph clearly shows that the area of density estimation spread shows the data coincides with each other. In general, this lexical type demonstrates some discriminative information and has a level of discriminative ability on the Twitter platform.

- Syntactic Features

To establish whether syntactic features have discriminative ability or not, further investigation of the Twitter platform has been conducted in order to demonstrate a comprehensive picture and to determine the top syntactic features used on this platform. Syntactic features have been ranked second out of the top most important features on the Twitter platform, specifically, the feature *number of punctuation marks,* as shown in Table 4-4. It is possible to establish the difference to which input data is similar or dissimilar between authors, and this feature improved the results for performance on Twitter, therefore it was repeated four times. As illustrated previously in Table 4-5, the lowest user EER was (0) *for Author 30,* while in contrast, the highest user EER was (38.14%) for *Author 22*. By performing density estimation examinations to determine the degree to which input data is similar or dissimilar between authors, this can be used for feature distribution because it has the ability to discriminate between authors. Figure 4-12 shows the plots of a univariate kernel density estimation to determine the differences in distribution for these top most important syntactic features between authors. The density estimates show the degree of discrimination for the *population*, lowest user EER (*Author 30)* and highest user EER (*Author 22)*.

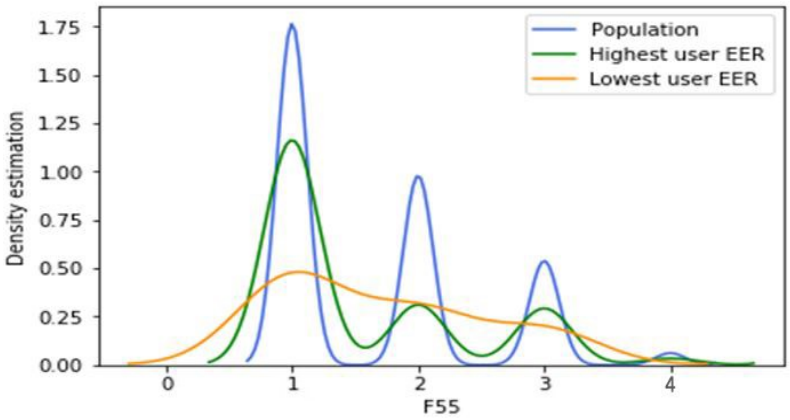


Figure 4-12: Density estimation plot for syntactic features (#punctuation)

In the above plots, the similarities and differences in input data between the authors can be seen. It shows that the *# punctuation* feature in syntactic has the potential to be used for discriminating users. It can be noticed that the *population* shares a similar feature vector with *Author 30* and *Author 22* with the *# punctuation feature*. However, it can be seen that *Author 30's* data is not similar, and there is a slight difference to *Author 22* and the *population*. This simple difference distinguishes *Author 30* from *Author 22* and the *population,* and shows that they are different. This because the *Author 22* feature vector centers on 1, 2 and 3, and their input vectors are not similar, or are not within entirely the same boundaries as the rest of the *population*. Thus, this is one of the most important syntactic features, which is what led the researcher to further illustrate how syntactic features have some level of discriminative ability for some authors.

Indeed, the analysis based on individual punctuation usage indicates that authors use common punctuation such as full stops, exclamation marks, colons, question marks and commas in Twitter messages, and it is possible to create a profile for punctuation marks on the Twitter platform because each person has their own punctuation style.

In general, syntactic features have demonstrated some discriminative information and have some level of discriminative ability that has led to improving performance for the Twitter platform.

- Structure Features

  In order to continue investigating the way a user organises the layout of messages posted between authors, an analysis of the discrimination of structural features on the Twitter platform has been conducted. The

purpose of this is to demonstrate a comprehensive picture to determine the top structure features used on this platform, as they have been ranked as the third most important features after lexical and syntactic, specifically the feature *total number of sentences,* and it has been shown that this feature is in the top 30 for the Twitter platform, as illustrated in Table 4-4. Furthermore, it is possible to establish the difference to which input data is similar or dissimilar between authors. As previously presented in Table 4-5, the lowest user error rate was (0) for *Author 30*, while in contrast, the highest user error rate was (38.14) for *Author 22*. Therefore, the investigation into the top structure feature is illustrated as *# sentences F209*. As expected, these features provide useful information for discrimination. For example, it is easy to discriminate *Author 30* from *Author 22* and from the *population*.

shows the plots of a univariate kernel density estimation to determine the differences in feature distribution for the most importance distribution structure features used to discriminate between the *population*; lowest user EER (*Author 30)* and highest user EER (*Author 22)*.

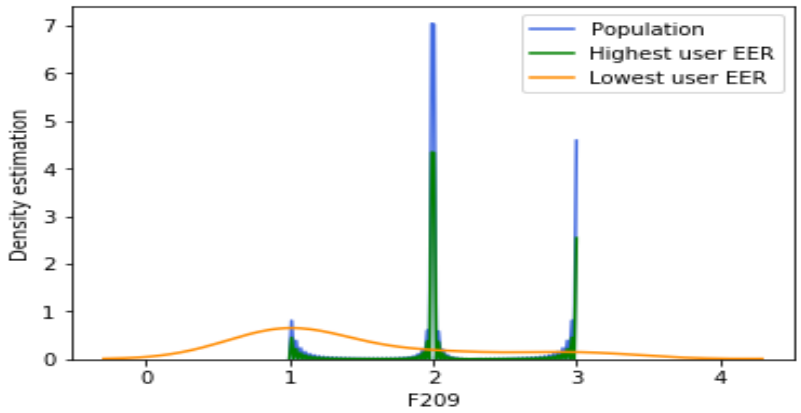



Figure 4-13 : Structure feature between population on Twitter (#Sentences)

From the above plots, the similarity and differences in input data between the authors can be seen. This feature may be useful for Twitter since it distinguishes *Author 30* from the *population*. It can be noticed that #sentences for the *population* centred around two to three sentences, while *Author 30* differs in that they centre around one sentence in their tweets. All in all, the #sentences of structure feature provides a robust feature that gives discriminative information that has a level of discriminative ability for the Twitter platform.

To conclude, with respect to the research question concerning Twitter recognition across the *population*, the performance of Twitter showed one of the worst performances, with an EER of 20.16% compared to other platforms, although it was better and outperformed the Facebook platform as that has an EER of 25% (Facebook will be explored later). This is because it is often used for public purposes and the writing style of authors is varied and different topics are discussed by different people, which may make it difficult to achieve high performance.

With respect to the investigation into the feature vector and how it impacts performance, it has been shown in Table 4-4 that lexical and syntactic features are the most repeated features and play an important role in discriminating between the *population*, thus improving the performance of Twitter.

#### 4.1.4.2 Text message platform

It may be beneficial to define the most discriminative population-based features among Text message users. In order to find the top most discriminating features for the Text message platform, 2,808 tests as a total experiment, and 106,359 samples of historical data of 26 authors, were conducted, as shown in Table 4-1.

By using the GB classifier, it was possible to correctly classify with an EER of 7.97% for the top 100 discriminating features.

### ➢ Top Most Important Features

In Text message experiments, the most discriminative features have been determined from the top 100 features of the best performance results, yielding an EER of 7.97% as shown in Table 4-2. The GB classifier outperformed other classifiers with data splitting 70/30 for training/ testing. A summary of the top features crossing EER is described in Figure 4-14 below.

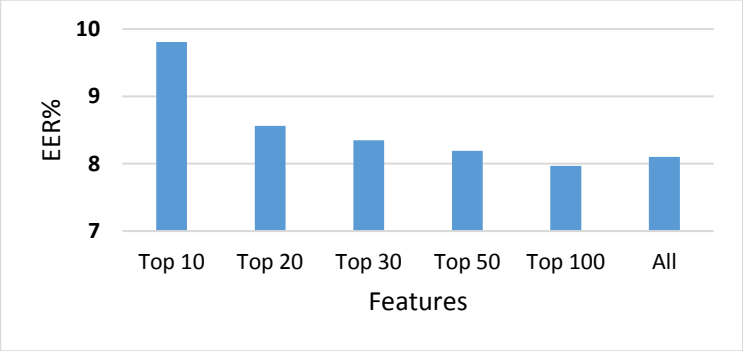



Figure 4-14: Top features with EER for the Text message platform

Table 5-6 shows the top 30 discriminating features analysed to show the similarities and differences in feature vectors between them and to focus on answering the research question: What commonalities and differences exist between the feature set for population?, by investigating the feature vector and how it impacts on performance, and exploring what commonalities and differences exist between authors on the Text message platform(a full listing of all 275 features can be found in the Appendix-B).

Table 4-6: Top Discriminative Features for population in Text messages

| Twitter Top discriminative Features in Twitter | | | |
|---|---|---|---|
| **No** | **Top 30 Features** | **Features category** | **Feature name** |
| 1 | 27 | Lexical | Number of alphabet a-z. |
| 2 | 232 | Short Messages feature | Frequency of missing a period or other punctuation to end a sentence. |
| 3 | 231 | Short Messages feature | Frequency of missing an uppercase letter when starting a sentence. |
| 4 | 52 | Syntactic | Number of punctuation. |
| 5 | 209 | Structure | Total number of sentences. |
| 6 | 215 | Lexical | Number of words with 1 char. |
| 7 | 233 | Short Messages feature | Frequency of missing the word “I” or “We” when starting a sentence. |
| 8 | 274 | Emotional feature | ❤ heart. |
| 9 | 1 | Lexical | Number of characters. |
| 10 | 3 | Lexical | Number of uppercase characters. |
| 11 | 228 | Short Messages features | Frequency of a smile face. |
| 12 | 51 | Syntactic | Number of punctuation. |
| 13 | 53 | Syntactic | Number of punctuation. |
| 14 | 2 | Lexical | Number of alphabets. |
| 15 | 54 | Syntactic | Number of punctuation. |
| 16 | 210 | Lexical | Total number of words. |
| 17 | 55 | Syntactic | Number of punctuation. |
| 18 | 213 | Lexical | Average sentence length in terms of character. |
| 19 | 214 | Lexical | Average sentence length in terms of word. |
| 20 | 211 | Lexical | Total number of short words (less than four characters). |
| 21 | 236 | Emotional feature | 😂 Face With Tears of Joy. |
| 22 | 212 | Lexical | Average word length. |
| 23 | 23 | Lexical | Number of alphabet a-z. |
| 24 | 12 | Lexical | Number of alphabet a-z. |
| 25 | 36 | Lexical | Number of special character. |
| 26 | 107 | Syntactic | Function words. |
| 27 | 58 | Syntactic | number of punctuation. |
| 28 | 56 | Syntactic | number of punctuation. |
| 29 | 217 | Lexical | Number of words with 3 chars. |
| 30 | 8 | Lexical | Number of alphabet a-z. |

| (Lexical) Top Repeated | (Syntactic ) Second Repeated | short message ) Third Repeated |
|---|---|---|

As demonstrated in Table 4-6, the most repeated features used when the top 30 features were captured are as follows: Lexical features were repeated over 15 times; followed by syntactic features repeated eight times; short message

features repeated four times, and finally, emotional features twice. Among the lexical features, it can be noted that the feature number of alphabet a-z category was repeated four times and also appeared at the top of the list. As can been noted, the type lexical was repeated and covered almost half of the dataset when the top 30 features were examined. The most distinguishing feature concerning lexical type was the category number of alphabet a-z, while in second place came the feature of syntactic, especially the category number of punctuation marks. While in third place came short message features and the category frequency of missing a period or punctuation to end a sentence or missing the word “I” or “We”, or missing an uppercase letter when starting a sentence. Furthermore, as expected, social emotion features may play a role on the Text message platform for author verification.

Therefore, number of alphabet a-z for lexical, and number of punctuation marks for syntactic, may be the most distinctive on the Text message platform when the first top thirty features are established, and this positively helps to improve performance.

These features have been used with the population to answer the part of research question on what commonalities and differences exist within the feature set for a population using Text message. Therefore, user level performance and feature distribution are discussed in the next section.

An analysis of the most discriminating features for a population needs to investigate user level and should, fundamentally, focus on two particular biometric characteristics: their capability to be universal and uniqueness, as shown in Figure 4-3 and Figure 4-4 for the descriptive statistical analysis inter-class and intra-class variance for some features. In order to select an effective and universal subset of features for individual platforms that can aid in author

verification, several stylometric features and social network specific features have been employed in this study. From a biometric perspective, with respect to universal (i.e. the category: number of alphabet a-z (lexical) repeated four times throughout the population), the feature of lexical is repeated over 15 times more than all other categories contained in the Top 30 features of the Text message platform; therefore, this has played a significant role in improving the performance of the Text message platform, as shown in Table 4-6.

Secondly, uniqueness was addressed, that is, determining which robust features categorise this platform and can be used for discriminating users. With the aim of presenting some useful insights into the identification of author features, and establishing the extent to which input data is similar or dissimilar between authors, it is important to investigate the ability to determine potentially positive features for establishing unique patterns to distinguish individual authors. There are two types of groups of authors concerning performance: for the first type, the authors have good performance, for example, the best case of individual performance achieved an EER of 0%. In contrast, the other group of authors showed poor performance, for example, the worst case of individual performance achieved an EER of 16.28%; therefore, the top 100 features are powerful for discriminating among populations. In general, poor user performance is due to author input vectors features, such as the use of different styles of writing making it difficult for the classifier. The analysis of the features shows the first top thirty features for each platform for analysis, and to illustrate that, the top features have some level of discriminative ability for some users - the users' features show the user distribution performance for *population, highest user EER* and *lowest user EER*. It can be seen that there are differences between the highest user's EER *(Author*

*9),* and the lowest user's EER (*Author 27)*. The analysis of feature vector distribution between users is presented in the next section.

Table 4-7: Users' EER for Text messages

| Users | EER | Users | EER |
|---|---|---|---|
| **1** | 5.57 | **14** | 13.24 |
| **2** | 5.79 | **15** | 0 |
| **3** | 0 | **16** | 5.41 |
| **4** | 13.29 | **17** | 14.24 |
| **5** | 6.25 | **18** | 7.32 |
| **6** | 12.24 | **20** | 4.56 |
| **7** | 11.47 | **24** | 14.14 |
| **8** | 12.48 | **25** | 8.18 |
| **9** | 16.28 | **27** | 0 |
| **10** | 13.01 | **30** | 3.91 |
| **11** | 4.82 | **38** | 4.25 |
| **12** | 6.41 | **39** | 13.56 |
| **13** | 3.57 | **40** | 7.21 |

| | |
|---|---|
| | Highest User EER |
| | Lowest User EER |

In order to investigate what commonalities and differences exist within the feature set for the population using the Text message platform, the top most important categories have been analysed to see how they contribute towards discrimination between the population as well as enhance performance. Since the analysis of the top six categories is often repeated when analysing within the features itself, the top thirty features, including their categories, were analysed. Therefore, the first top thirty features, including their categories, are presented next.

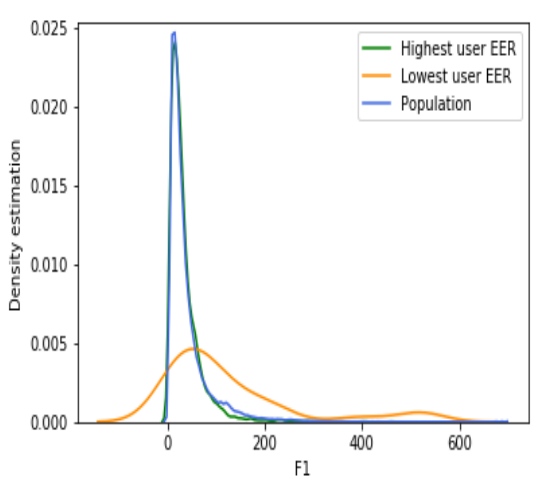
### ➢ Examining the Top Thirty Important Features

#### • Lexical Features

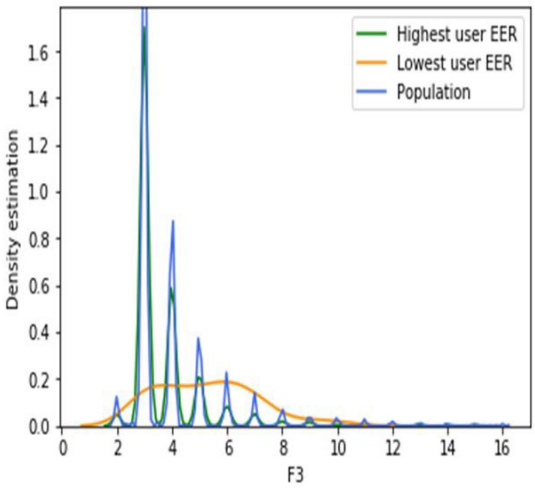
This section examines the way that these types of features have an impact on the performance of Text messages and establishes the extent to which input data is similar or dissimilar between authors. As illustrated previously, the top feature for the Text message platform is the lexical

feature *total number of alphabet a-z,* as this has the ability to discriminate between authors and is active on this platform since it has discriminative information, therefore it was repeated, as shown in Table 5-6. Further investigation into other top lexical features on the Text message platform has been conducted to provide a comprehensive picture and to determine the top lexical features utilised on this platform, as lexical features have been ranked top for the Text message platform when population-based features are performed. It is possible to establish the difference to which input data is similar or dissimilar between authors on the Text message platform. As shown earlier in Table 4-7, the lowest user error rate was (0) for *Author 15,* while in contrast, the highest user error rate was (16.28) for *Author 9*. Therefore, an investigation into these top lexical features has been conducted and number of characters and number of uppercase characters, F1 and F3 respectively, have been investigated, since they are the top most important lexical features from amongst other lexical features for the Text message platform. By performing density estimations, it is possible to establish the degree to which input data is similar or dissimilar between authors, and this can be used for feature distribution.

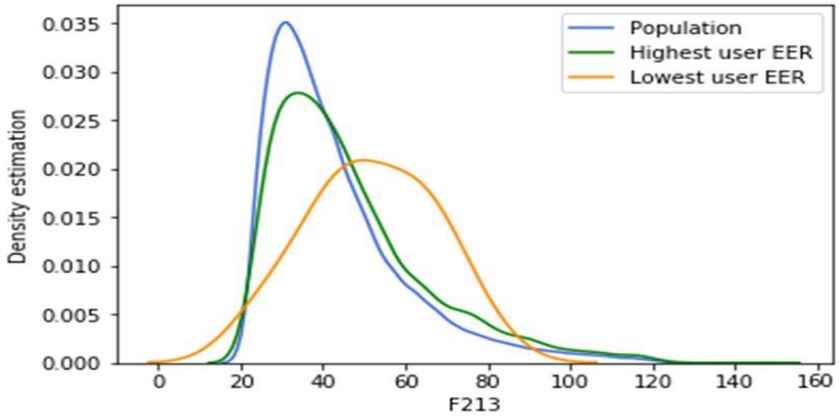
Figure 4-15 shows plots of a univariate kernel density estimation to determine the differences in feature distribution for these lexical features between authors. The density estimates show the degree of discrimination for the *population*, lowest user EER (*Author 15)* and highest user EER (*Author 9)*.




(a) # Character




(b)# Uppercase characters




(c) Average sentence length in terms of character

Figure 4-15: Top lexical distribution features between population in Text messages

The above plots show that the lexical features *# character*, *# uppercase characters and Average sentence length in terms of characters,* have the potential to be used to discriminate between users. The similarities and differences in input data between the authors can be seen. In plot (a) *# character F1*, it can be noticed that the population of authors do not share a similar feature vector, with *Author15*, showing only a slight similarity between them. Most of the *population* and *Author 9* have centered around 40 characters in their Text messages, while *Author 15* differs and has centered around 60, as well as centering on 500 characters.

On the other hand, it can be seen that *Author 15* and *Author 9* are not similar, and this simple difference distinguishes them from the *population*, as well as distinguishing them from each other due to their input vectors not being similar and not within entire boundaries compared to the overall *population*. Therefore, this feature is one of the most important popular features from the lexical category since some authors can be seen to use characters that are different from others. Indeed, the number of characters feature plays an important role in the Text message platform for the verification process. Therefore, this feature makes it effective and active on the Text message platform for verification since it discriminates, for instance, *Author 15* and the *population* have been discriminated from each other, as shown on the graph.

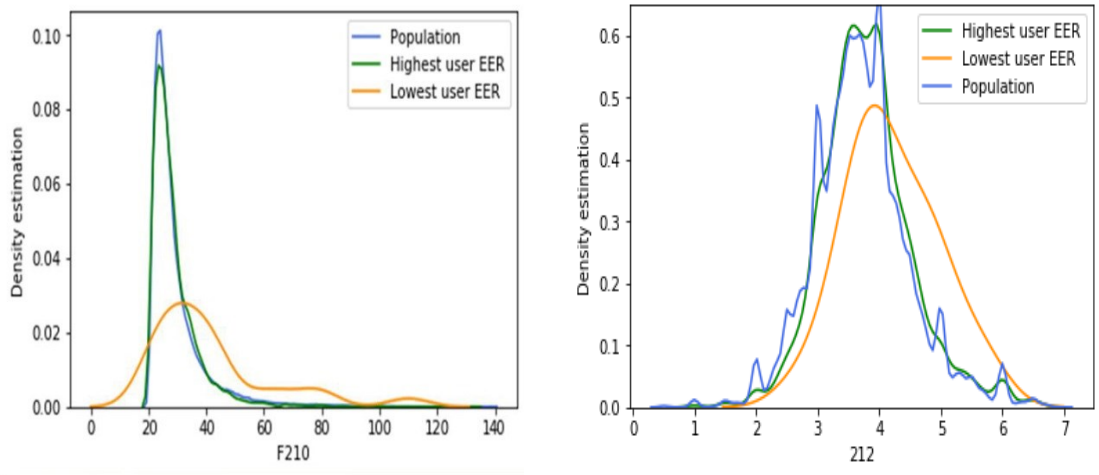
Moving to another important lexical feature (b) *# Uppercase characters F3.* It can be seen that the feature vector area does not coincide between *Author15* and the *population* and *Author 9*. Indeed, this explains the nature of this feature in that it varies from user to user and plays a role in the discrimination process between authors. Therefore, this feature is effective and active on this platform since it has shown some discrimination between *Author 15*, Author 9 and the *population,* as the graph clearly shows that the area plots of data do not coincide with each other.

The last important lexical feature on the Text message platform is *average sentence length in terms of character F213*, as shown in plot (c): *Author 15* uses this feature and feature vector as they have concentrated on around 60 characters for average sentence length, which means that *Author 15* often uses this number of characters in sentences on the Text message platform, which is different to other authors. However, it can be noted that *Author 9* is difficult to discriminate from the *population* as they clearly share a similar boundary with the *population* who use between one and 80 characters. This indicates that as long

as this feature creates discrimination, it is effective and active on this platform for the verification process, as it is possible for *Author 15* and *Author 9* to be discriminated from each other because the graph shows that the area plots of data do not coincide with each other or the population.

As a result, this feature, # number of alphabet a-z, #Characters, # Uppercase characters and average sentence length in terms of characters, has the ability to discriminate *Author 15* from *Author 22* and the *population* and provides a robust feature for the lexical category. Overall, lexical character-based features have some level of discriminative ability on the Text message platform.

To examine the way in which authors use short or long words in their Text message messages on the Text message platform (this feature is explored more in Section 6.5), the distribution of word number and average word length for population, user lowest error, and user highest user EER, were examined. Figure 4-16 below illustrates examples of the number of word distribution usage




a) Total # words (b) Average word length

Figure 4-16: Density estimation plot for total number of words and average word length features

An analysis of the number of word feature indicates that the majority of authors on Text message used an average of two to 40 words in their text messages.

However as can be seen in Figure 4-16, most authors tend to use short words on the Text message platform, and centre around approximately 25 words. This is expected, as authors have to find a way of being concise and short in their Text messages with a limited number of words. On the other hand, it can also be noted that *Author 15* uses more words, centred around 17, than *population,* and so this feature definitely distinguishes that user from others. Therefore, this is a robust feature that provides discriminative information which has a level of discriminative ability for the Text message platform.

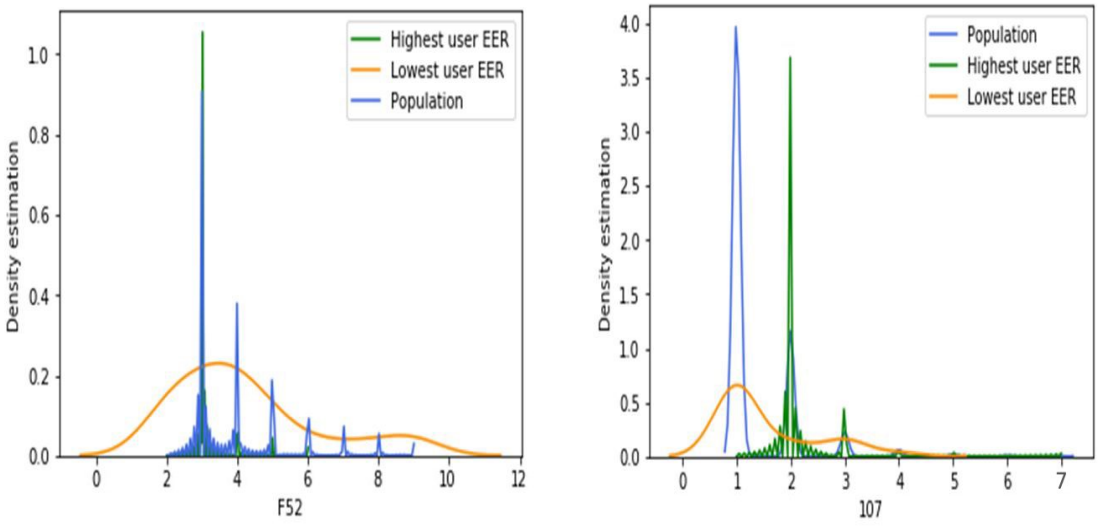
Further analysis has been conducted, and plot (b) describes the average word length distribution usage for *population*, *lowest user EER (i.e.15)* and *highest user EER (i.e. 9)*. It can be seen that most of the *population* tend to use an average number of words in their Text messages that are four characters long. However, it is difficult to distinguish *Author 15*. This is expected, because they are restricted to a certain number of characters on the Text message platform and authors tend to try to be concise in their SMS text messages. All in all, as was expected, this measure is not a robust feature because *Author 15* cannot be discriminated from the *population,* which indicates that input vectors are more likely to be similar between authors, and the graph clearly shows that the data for the area of the authors' density estimation coincides with each other.

- Syntactic Features

  To establish whether syntactic features have discriminative ability or not, as well as whether they have an effective impact on the results for the Text message platform, further investigation has been conducted. This has provided a comprehensive picture and addresses the top most syntactic features used on this platform, which has been ranked second from the top

for features, specifically, the feature *number of punctuation marks,* as shown in Table 4-6.

It is possible to establish the difference to which input data is similar or dissimilar between authors. As illustrated previously in Table 4-7**,** the lowest user error rate was (0) for *Author 15,* while in contrast, the highest user error rate was 16.28% for *Author 9*. The top most important categories amongst other syntactic features on the Text message platform are: *Number of punctuation marks F52* and *#function words F107*. By performing density estimation examinations, it is possible to determine the degree to which input data is similar or dissimilar between authors, which can be used for feature distribution because it has the ability to discriminate between authors. Figure 4-17 shows the plots of a univariate kernel density estimation to determine the differences in features distribution for the top most importance distribution syntactic features between authors. Density estimating shows the degree of discrimination for *population*, lowest user EER (*Author 15)* and highest user EER (*Author 9)*.




(a) #punctuation. (b) #function words.

Figure 4-17: Density estimation plot for syntactic features in Text messages

In the above plots, the similarities and differences in input data between the authors can be seen, and it is clear that a number of punctuation features have the potential to be used for discriminating users. It can also be noticed that the *population* does not share similarities in density distribution, for example *Author 15* and *Author 9* with (a)*# punctuation feature* and (b)*# function words*. In addition, it can be seen that the density estimation area plot of data does not coincide between *Author15* and the population and *Author 9*. Most of the *population* and *Author 9* do use it in a constant manner that centers around a specific number, while *Author 15* is different and has no specific centration. Indeed, this explains the nature of this feature in that it varies from user to user and has a role in the discrimination process between authors. Therefore, this feature is effective and active on this platform since it has shown some discrimination between authors, for instance, *Author 15*, *Author* 9 and the *population* can be discriminated from each other, as the graph clearly shows that the area plots of data do not coincide between each other or with the *population*.

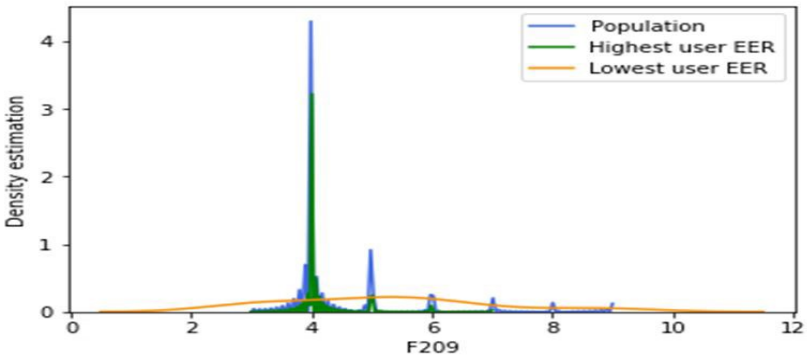
In general, some syntactic features demonstrate some discriminative information and have some level of discriminative ability on the Text message platform, and this has positively helped to improve the results of its performance in SMS Text messages.

- Structure Features

  This section will investigate the way the layout of Text messages is organised between authors. An analysis of the discriminating structural features on the Text message platform is provided, with the purpose of providing a comprehensive picture to determine the top structure features used on this platform. Structure has been ranked one of the top most important features on the Text message platform, especially the feature

Total # Sentences, as shown in Table 4-6. It is possible to establish the difference to which input data is similar or dissimilar between authors. As previously presented in Table 4-7, the lowest user error rate was (0) for *Authors 15,* while in contrast, the highest user error rate was (16.28) for *Author 9*. Therefore, an investigation into the top structure feature has been conducted: Total number of sentences, F209, have been analysed, since it has been shown that this is the most important structure from among other structure features in the top 30 for the Text message platform.

As expected, these features provide some useful information for discrimination. For example, whether it is easy to discriminate between *Author 15, Author 9* and the *population*. By performing density estimation examinations, the degree to which input data is similar or dissimilar between populations was established, and this can be used for feature distribution because it has the ability to discriminate between authors. Figure 4-18 shows the plots of a univariate kernel density estimation to determine the differences in feature distribution for these top distribution structure features, in particular, lowest user EER (*Author 15)* and highest user EER (*Author 9)*.




# Sentences

Figure 4-18: Density estimation plot for structure features on the Text message platform

From the above plots, the similarities and differences in input data between the authors and population can be seen. The graph shows that structure features

have the potential to be used for discriminating users from Text messages. In the above plot, it can be seen that *population* does not share much of a similar feature vector with *Author 15* for *# sentences*. Most of the *population* centered on around two, three or four sentences, while *Author 15* fluctuates, which distinguishes this author from the *population*.

- Emotional Feature Distribution in Text message Platform

  Emotional feature is one of the most important features on the Text message platform, as shown in Table 4-6. Amongst all other emotional categories, the one ranked top and an effective discriminative feature is the Text message emotional symbol ❤ *(heart), F274*. This suggests that it is popular and commonly utilised between users on the Text message platform.

In order to examine the similarity of feature vectors based on the way a user uses this emotional icon in their Text messages, the density distribution/pattern for this feature has been estimated, as shown in Figure 4-19 below.

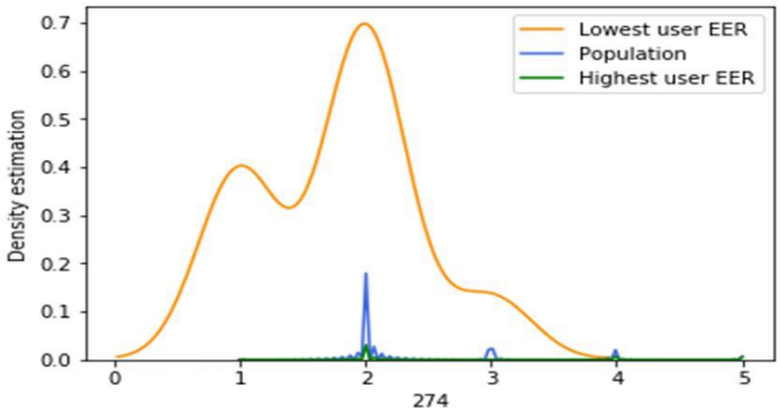



Figure 4-19: Emotional feature on the Text message platform

Figure 4-19 shows that individual authors, for example for the lowest EER (*Author 15),* can be discriminated from the *population,* as it can be seen that the density distribution/pattern of this feature for *Author 15* distinguishes the author from

others, as their plotting area does not coincide. The density distribution/pattern for the lowest user rate (*Author 9)* and the *population* is not shared to the same extent as *Author 15,* as shown in Figure 4-19. This suggests that some users are using this feature more and some are not. The intensity of the use of this feature certainly distinguishes users of this feature.

All in all, emotional feature analysis shows that the way emotional features are used contributes to providing additional discriminative information and thus improves performance. Therefore, as expected, this feature can contribute to providing a robust feature on the performance of the Text message platform, which is an important point as no previous research has explored the use of emoticons in Text message platform (to the best of the author's knowledge).

To conclude, with respect to the research question on the performance of Text message recognition across the population, one of the best performances was an EER of 7.97% compared to other platforms, and this outperformed all other platforms. This result is likely to be because for Text messages the writing capacity is often small, and it is considered a private platform - often one to one - unlike Twitter. Although Twitter has small capacity, it is considered a public platform for one to many users, which could make it more reliable for achieving high performance.

With respect to investigating the feature vector and how it impacts performance, it has been found that the lexical feature was repeated more (e.g. fifteen times), next was syntactic (e.g. four times), while other features such as emoticon and short message features could also play an important role in discriminating between the population and thus improving performance on the Text message platform.

#### 4.1.4.3 Facebook Platform

It is beneficial to define the most discriminative population-based features for Facebook users. Therefore, in order to find the top most discriminating features for the Facebook platform, 4,968 tests as total experiments were conducted, and 4.539 samples of historical data from 47 authors included, previously shown in Table 4-1. By using the GB classifier, it was possible to correctly classify the top most discriminating features and the best performance achieved, which is an EER of 25%, with approximately 75% accuracy rate.

- **Top Most Important Features**

  In the Facebook experiments, the top and most discriminative features have been determined from all 275 features employed, as shown previously in Table 4-2. Three classifiers were utilised: SVM, GB and RF. GB was the best and most appropriate classifier, with data split 70/30 for training/ testing. A summary of the crossing of the top features and their accuracy is shown in Figure 4-20 below.

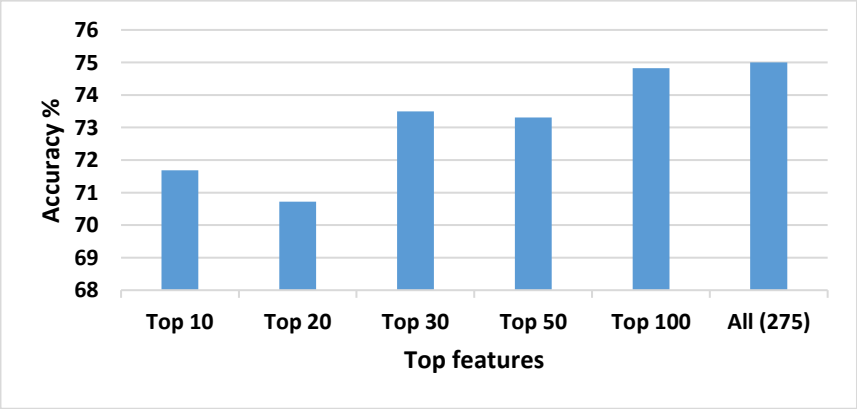



Figure 4-20: Top features and their accuracy on the Facebook platform

Table 4-8 below shows the top 30 features on the Facebook platform. In order to answer the research question: What commonalities and differences exist within the feature set for population? The feature vector and how it impacts performance

was investigated, as well as what commonalities and differences exist within the feature set for the population; in addition, an analysis of the similarities and differences in feature vectors between authors on the Facebook platform is provided. The procedures of analysis will be the first Top 30 features.

Table 4-8: Top Discriminative Features among the population for Facebook

| Top discriminative Features in Facebook | | | |
|---|---|---|---|
| **No** | **Top 30 Features** | **Features category** | **Feature name** |
| 1 | 52 | Syntactic | Number of punctuation. |
| 2 | 55 | Syntactic | Number of punctuation. |
| 3 | 54 | Syntactic | Number of punctuation. |
| 4 | 1 | Lexical | Number of characters. |
| 5 | 2 | Lexical | Number of alphabets. |
| 6 | 231 | Short Message feature | Frequency of missing an uppercase letter when starting a sentence. |
| 7 | 213 | Lexical | Average sentence length in terms of character. |
| 8 | 212 | Lexical | Average word length. |
| 9 | 3 | Lexical | Number of uppercase characters. |
| 10 | 214 | Lexical | Average sentence length in terms of word. |
| 11 | 210 | Lexical | Total number of words. |
| 12 | 209 | Structure | Total number of sentences. |
| 13 | 32 | Lexical | Number of special character. |
| 14 | 232 | Short Message feature | Frequency of missing a period or other punctuation to end a sentence. |
| 15 | 8 | Lexical | Number of alphabet a-z. |
| 16 | 22 | Lexical | (Number of alphabet a-z. |
| 17 | 48 | Lexical | Number of special character. |
| 18 | 23 | Lexical | Number of alphabet a-z. |
| 19 | 12 | Lexical | Number of alphabet a-z. |
| 20 | 228 | Short Message feature | Frequency of a smile face. |
| 21 | 233 | Short Message feature | Frequency of missing the word “I” or “We” when starting a sentence. |
| 22 | 211 | Lexical | Total number of short words (less than four characters). |
| 23 | 58 | Syntactic | Number of punctuation. |
| 24 | 11 | Lexical | Number of alphabet a-z. |
| 25 | 216 | Lexical | Number of words with 2 chars. |
| 26 | 236 | Emotional feature | 😂 Face With Tears of Joy. |
| 27 | 17 | Lexical | Number of alphabet a-z. |
| 28 | 4 | Lexical | Number of alphabet a-z. |
| 29 | 53 | Syntactic | Number of punctuation. |
| 30 | 24 | Lexical | Number of alphabet a-z. |

| (Lexical) Top Repeated | (Syntactic ) Second Repeated | ( Short Messages) Third Repeated |
|---|---|---|

As shown in Table 4-8, the most repeated features used when the top thirty of the feature vector was captured are as follows: Lexical features were repeated over 18 times, followed by syntactic features were repeated five times, and finally, short message features were repeated three times. While structure and

emoticons appeared once for each one. In general, lexical and syntactic features are the most discriminating on the Facebook platform.

To investigate the effect of the top most discriminating features among the population on the Facebook platform, an experimental analysis of features for population has been conducted to provide a comprehensive picture and gain a better understanding of the nature of these discriminative features as an input vector for population. These have been ranked as the most significant discriminative features for the population on the Facebook platform. User level performance and feature distribution are provided in the next section.

To investigate the most discriminating features between authors on the Facebook platform, an experimental analysis of features between authors is provided to demonstrate a comprehensive picture of these top 275 most discriminative features as the input vector among users, as shown previously in Table 4-8.

From a biometric perspective, with respect to being universal, the number of punctuation marks category (syntactic) appears five times throughout population, while lexical appears 14 times, which is more than all other types for the Facebook platform. Therefore, these syntactic and lexical features play a significant role in improving the performance of the Facebook platform, as shown above in Table 4-8.

Secondly, with respect to uniqueness, the most robust features that categorise this platform and can be used for discriminating users have been explored. The aim of this is to present some useful insights into the identity of author features. In addition, it is necessary to establish the differences in input data, and whether it is similar or dissimilar between authors, therefore it is important to investigate the ability to determine potentially positive features for establishing unique

patterns to distinguish individual authors. The best case of individual performance achieved an EER of 5%. In contrast, the other group of authors showed poor performance, for example the worst case of individual performance achieved an EER of 47.21%; therefore, the top 275 features are powerful for discriminating between populations. In general, poor user performance due to author input vector features includes the use of different styles of writing, making it difficult for the classifier.

Table 4-9 shows the distribution of the top 275 features for *population,* the *highest user EER* and the *lowest user EER*. It can be noticed that there are differences between the highest user's EER (47.21%) *(Author 48),* and the lowest user's EER (5%) (*Author 34)*, and an analysis of the most important features has been conducted. Therefore, a series of investigations has been conducted, and the analysis of feature vector distribution between users is presented in the next section, which illustrates that the top features have some level of discriminative ability.

Table 4-9: Users' EER On Facebook

| Users | EER | Users | EER |
|---|---|---|---|
| **1** | 20.86 | **24** | 12.5 |
| **2** | 21.87 | **25** | 26.13 |
| **3** | 18.05 | **26** | 21.82 |
| **4** | 6.25 | **27** | 33.86 |
| **5** | 25.45 | **28** | 23.61 |
| **6** | 27.74 | **29** | 19.37 |
| **7** | 27.61 | **31** | 20.70 |
| **8** | 29.90 | **32** | 27.34 |
| **9** | 30.60 | **33** | 23.61 |
| **10** | 34.05 | **34** | 5 |
| **11** | 21.18 | **35** | 23.98 |
| **12** | 18.74 | **36** | 36.80 |
| **13** | 28.75 | **37** | 36.66 |
| **14** | 37.79 | **38** | 12.23 |
| **15** | 29.11 | **39** | 6.25 |
| **16** | 32.79 | **40** | 40.66 |
| **17** | 9.610 | **41** | 28.70 |
| **18** | 18.69 | **42** | 34.62 |
| **19** | 24.92 | **43** | 18.33 |
| **20** | 27.61 | **46** | 40.83 |
| **21** | 16.29 | **47** | 24.94 |
| **22** | 26.13 | **48** | 47.21 |
| **23** | 20.32 | **50** | 30.24 |

| | |
|---|---|
| | **Highest User EER** |
| | **Lowest User EER** |

In order to investigate what commonalities and differences exist within the feature set for the population for the Facebook platform, the top most important features, including categories, have been analysed to investigate how each feature contributes towards discriminating between the population. In the next section, the first top thirty features and their categories are presented.

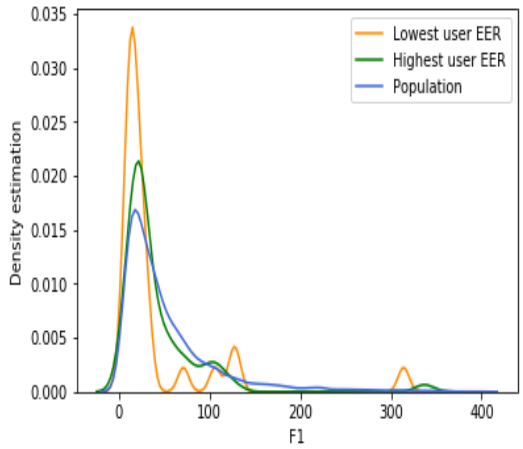
### ➢ Examining the top thirty most important features

#### • Lexical Features

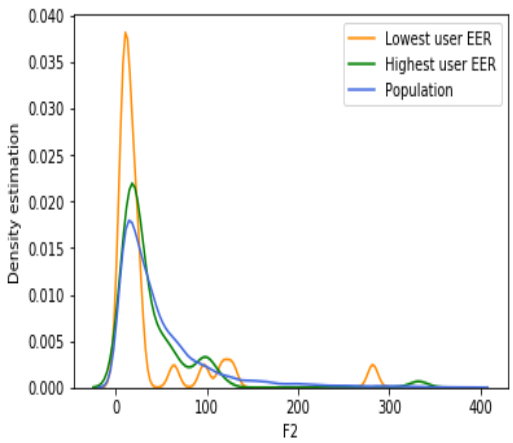
This section will examine the way in which these types of features have an impact on the performance of Facebook, and to establish which input data is similar or dissimilar between authors on the Facebook platform. Lexical

features are the top most important type of feature on Facebook, as shown in Table 4-8, and it is possible to establish the extent to which input data is similar or dissimilar between authors.

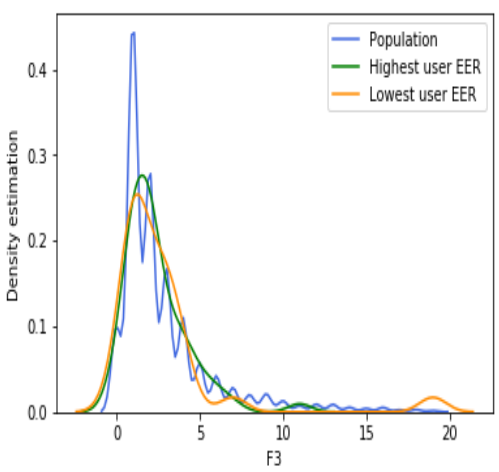
As shown previously in Table 4-9, the lowest user error rate was (5) for *Author 34,* while in contrast, the highest user error rate was (47.21) for *Author 48*. Therefore, lexical feature, number of characters (F1), number of alphabets (F2), average sentence length in terms of characters (F213), and number of uppercase characters (F3), have been investigated. This is because it was discovered that these are the most important lexical features from amongst other lexical features on the Facebook platform. By performing density estimations to establish the degree to which input data is similar or dissimilar between populations. Figure 4-21 shows the plots of a univariate kernel density estimation to determine the differences in feature distribution for these top most important lexical features used to discriminate between the population, and the lowest user EER (*Author 34)* and the highest user EER (*Author 48)*.

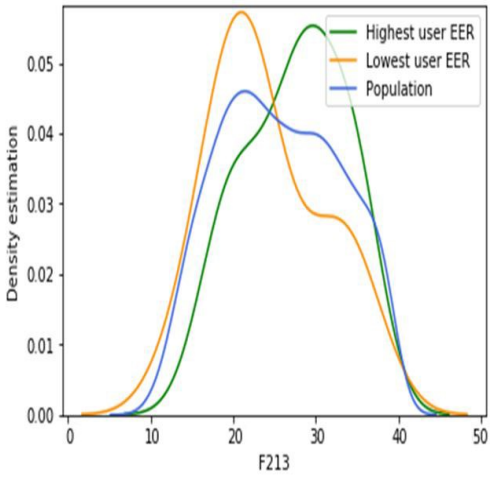



(a) # Characters




(b) # Alphabets



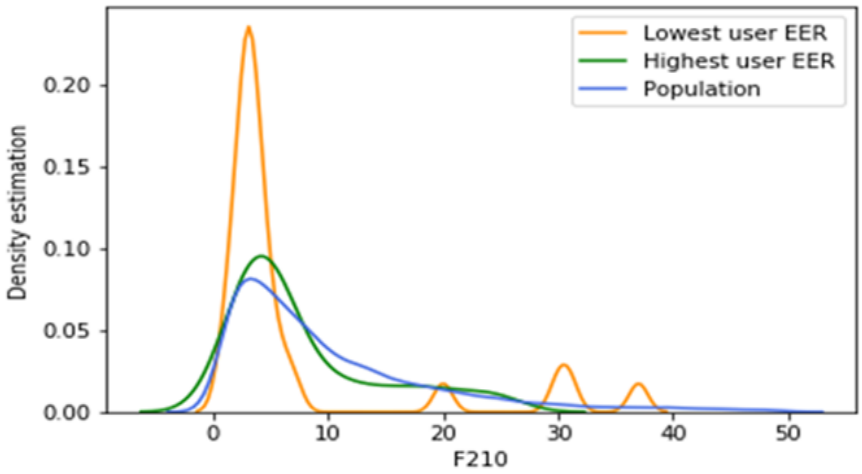




(c) # Uppercase characters (d) Average sentence length in terms of character

Figure 4-21: Top distribution lexical features between the population on the Facebook platform

Figure 4-21 shows that a number of features have the potential to be used for discriminating users. It shows the top lexical character based feature distribution; for example, *Number of characters* (F1), *number of uppercase characters* (F3), *number of alphabets* (F2), and *average sentence length in terms of characters* (F213), which have all been examined. Figure 4-21 presents the differences between the highest user EER, the lowest user EER and the population, providing an estimate of the similarity between authors' input vectors. First, in plot (a) #Characters F1, the lowest user error rate *(Author 34)* is different from *Author 48* and the rest of the *population*. Therefore, there is some level of discriminative ability between them. For example, it is possible for *Author 34* – the author with the lowest error (see Figure 4-21 (a) and (b) and (d)) - to be discriminated from others, as the graph clearly shows that the areas plot of data do not coincide with each other. On the other hand, it is difficult to discriminate *Author 34* from *Author 48* or the *population* (see Figure 4-21 (c)) as they cover the same area of data between each other. As a result, this feature is robust because it has the ability to distinguish between authors.

To investigate the way in which authors use short or long words in their posts on the Facebook platform (this feature is explored more in Section 6.5), the distribution of word number for *population*, *user lowest error*, user *highest user error* were investigated. Figure 4-22 illustrates examples of number of word distribution.




# Words

Figure 4-22: Number of words for authors

An analysis of the number of words feature indicates that the majority of authors on the Facebook platform used words that were an average of two to 40 words long in their posts. However, as can be noted in Figure 4-22, most authors tend to use short words in their posts of between two to 10 words. This is expected, as Facebook users have to find a way of being short in their social post messaging with a limited number of words (Hussain et al., 2014). Also notice that *Author 34*, used this feature more between two and 10 words, which makes them easy to identify and distinguish from this word limit compared to the *population*.

- Syntactic Features

  In order to continue investigating discriminative ability based on syntactic features and punctuation features on the Facebook platform, experimental

analysis has been conducted. It is possible to establish the differences in feature vectors and to what extent input data is similar or dissimilar among the population.

Table 4-8 above shows that a number of syntactic features have the potential to be used for discriminating users. Figure 4-23 shows the top syntactic based features, for example, *number of punctuation marks (F55)*, and *number of punctuation marks (F54)* were examined.

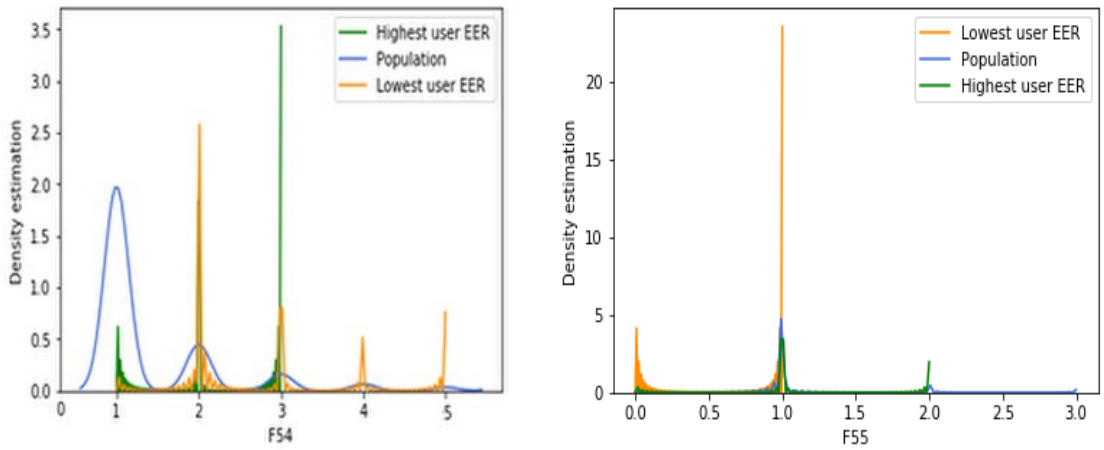



Figure 4-23: number of punctuation features between authors on the Facebook platform

As demonstrated in Figure 4-23, for the best cases - *Author 34* with low EER and *Author 48* with high EER- used the same punctuation marks in their message profiles, however, this feature can distinguishe them from others. The analysis is based on individual punctuation usage, and indicates that authors used common punctuation such as full stops, exclamation marks, colons, question marks and commas in post messages. It is possible to create a profile for punctuation marks on the Facebook platform because it is clear that each person has their own punctuation style. Therefore, this feature has some discriminative ability for Facebook because it separates authors' usage of punctuation on the Facebook platform.

- Structural Features

This section will investigate the way the layout of post messages is organised between authors. An analysis of discriminating structural features on the Facebook platform is provided, with the purpose of presenting a comprehensive picture to determine the most useful structural features used on this platform out of those ranked as the top features. Therefore it is possible to establish the difference to which input data is similar or dissimilar between authors.

As presented previously in Table 4-9, the lowest user error rate was (5) for *Author 34,* while in contrast, the highest user error rate was (47.21%) for *Author 48*. Therefore, the top structural feature on the Facebook platform (Total number of sentences, F209) has been investigated. It can be said that these features provide useful information for discrimination. For example, it is easy to discriminate *Author 34* from *Author 48* and the *population*. By performing density estimation examinations, the degree to which input data is similar or dissimilar between populations was established, which can be used for feature distribution because it has the ability to discriminate between authors. Figure 4-24 shows the plots of a univariate kernel density estimation to determine the differences in feature distribution for the most important distribution structure features used to discriminate the population. This shows the lowest user EER (*Author 34)* and the highest user EER (*Author 48)*.

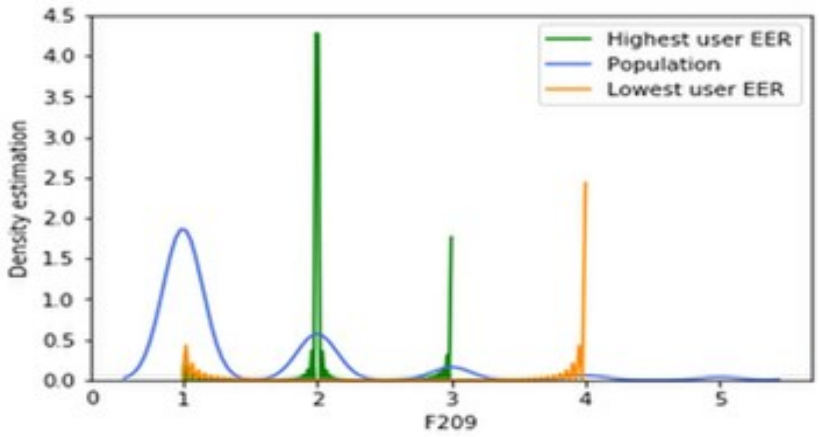



Figure 4-24 : Density estimation plot for the number of sentences on the Facebook platform

Further analysis was carried out on structural features, and Figure 4-24 shows the total number of sentences on the Facebook platform. As expected, the total number of sentences for the population on Facebook ranged from approximately one to four sentences and less, because the nature of Facebook is for posts to be short (Li et al., 2016). It can be noticed that the difference between *Author 34* and *Author 48 is as* follows: *Author 34* often focused on one or four sentences, while *Author 48* focused on two or three sentences, thus, distinguishing them from the *population* when using the feature number of sentences. Thus, this feature can provide discriminative ability on Facebook because it separates their usage. Please note that average sentence length in terms of characters and total number of words (F213 and F210 respectively) were discussed in the Lexical Features section.

To conclude, with respect to the research question concerning the performance of Facebook recognition across the population, the performance was worse, with an EER of 25%. This is probably because it is a public platform and message topics and subjects vary between authors.

With respect to investigating the feature vector and how it impacts performance, it can be seen in Table 4-8 that lexical features were repeated more (eighteen times), next were syntactic five times (five times), and so these play an important role in discriminating the population and thus improved performance for the Facebook platform.

#### 4.1.4.4 Email platform

It is beneficial to define the most discriminative population-based features for Email users. Therefore, in order to find the top most discriminating features for the Email platform, 5,076 tests as total experiments, and 6,540 samples of historical data from 47 authors were analysed, as shown previously in Table 4-1. By using the GB classifier, it was possible to correctly classify the top most discriminating features, and the best performance achieved was an EER of 13.11%.

- **Top Most Important Features**

  For Email experiments, the top and most effective features have been determined out of 100 features to find the best performance results, yielding an EER of 13.11%, as shown previously in Table 4-2. GB was the best and most appropriate classifier with data split 70/30 for training/ testing. A summary of crossing the top features with EER is shown in Figure 4-25 below.

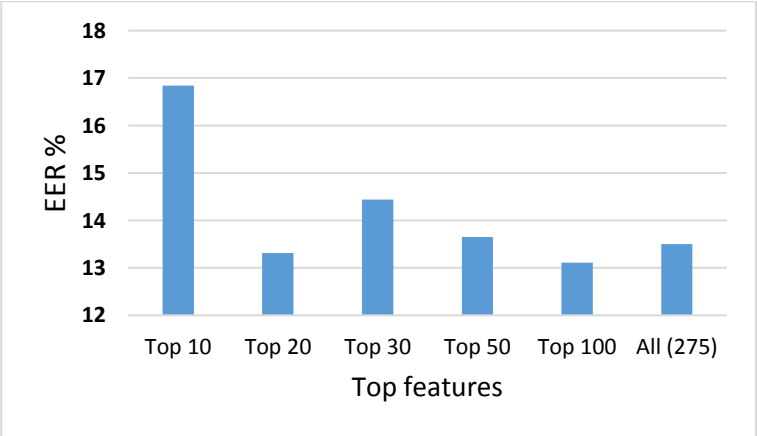



Figure 4-25: Top features and EER for the Email platform

Table 4-10 shows the top 30 features for the Email platform. In order to answer the research question: What commonalities and differences exist within the feature set for population?, the feature vector and how it impacts performance has been investigated. In addition, an analysis of the similarities and differences in feature vectors between authors for the Email platform is provided (a full listing of all 275 features is provided in the Appendix).

Table 4-10: Top Discriminative features among the population for Email

| Top discriminative Features in Email | | | |
|---|---|---|---|
| **No** | **Top 30 Features** | **Features category** | **Feature name** |
| 1 | 29 | Lexical | Number of alphabet a-z ("z"). |
| 2 | 38 | Lexical | Number of special character ("- "). |
| 3 | 55 | Syntactic | Number of punctuation (":"). |
| 4 | 50 | Syntactic | Number of punctuation ("\|"). |
| 5 | 39 | Lexical | Number of special character ("_"). |
| 6 | 102 | Syntactic | Function words ("from"). |
| 7 | 48 | Lexical | Number of special character ("/"). |
| 8 | 51 | Syntactic | Number of punctuation (","). |
| 9 | 52 | Syntactic | Number of punctuation (".") . |
| 10 | 231 | Short Message features | Frequency of missing an uppercase letter when starting a sentence. |
| 11 | 42 | Lexical | Number of special character (">"). |
| 12 | 228 | Short Message features | Frequency of a smile face ( ":)"). |
| 13 | 227 | Lexical | Number of words with more than 12 chars. |
| 14 | 43 | Lexical | Number of special character ("<"). |
| 15 | 212 | Lexical | Average word length. |
| 16 | 213 | Lexical | Average sentence length in terms of character. |
| 17 | 3 | Lexical | Number of uppercase characters. |
| 18 | 54 | Syntactic | Number of punctuation ("!"). |
| 19 | 13 | Lexical | Number of alphabet a-z ("j"). |
| 20 | 58 | Syntactic | Number of punctuation (" ' "). |
| 21 | 6 | Lexical | Number of alphabet a-z ("c"). |
| 22 | 14 | Lexical | Number of alphabet a-z ("k"). |
| 23 | 31 | Lexical | Number of special character("@"). |
| 24 | 126 | Syntactic | Function words ("my"). |
| 25 | 214 | Lexical | Average sentence length in terms of word. |
| 26 | 27 | Lexical | Number of alphabet a-z ("x"). |
| 27 | 56 | Syntactic | Number of punctuation (";"). |
| 28 | 26 | Lexical | Number of alphabet a-z ("w"). |
| 29 | 1 | Lexical | Number of characters. |
| 30 | 21 | Lexical | Number of alphabet a-z("r"). |

| (Lexical) Top Repeated | (Syntactic ) Second Repeated | ( Short Messages) Third Repeated |
|---|---|---|

Table 4-10 shows the type of lexical was repeated over 18 times, while the most repetitive lexical features are *# alphabet a-z* seven times (it was also the first feature on the list), and category *#special characters,* which was repeated five times. In second place came syntactic features, which were repeated nine times,

in particular, the feature number of *punctuation* marks was repeated six times, and the feature short messages appeared twice. It can be noticed that the feature structure on the Email platform disappeared. Therefore, number of special characters, number of alphabet a-z for lexical, and number of punctuation marks for syntactic seem to be the most distinctive features on the Email platform once the top thirty discriminating features have been established, and this positively helped improve performance. To investigate the effect of the most discriminating features on the Email platform, user level performance and feature distribution are discussed next.

To investigate these discriminating features between authors on the Email platform, an experimental analysis of features between authors has been conducted. These have been ranked the most significant discriminative features for the Email platform, as shown in Table 4-10 above. With respect to universal, lexical appears 18 times throughout the *population*, while syntactic appears eight times. Secondly, uniqueness has been considred, such as determining which robust features can be categorised and used for discriminating users. With the aim of presenting some useful insights into the identity of author features, establishing the extent to which input data is similar or dissimilar between authors is important, in order to investigate the ability to determine whether there are two types of groups for authors concerning their performance: In the first type, the authors showed good performance, for example, the best case of individual performance achieved an EER of 0%. In contrast, the other group of authors showed poor performance, for example, the worst case of individual performance achieved an EER of 27.77%, as shown in Table 4-11. In general, poor author performance was due to author input vector features that are similar and overlap, or are located within similar boundaries to other authors.

Table 4-11 shows the top features distribution for population, highest user EER and lowest user EER. The differences between the highest user EER (bright orange colour) for *Author 33*, and in contrast, the lowest user EER (blue colour) for *Author 22,* can be seen.

Table 4-11: Authors' EER for Email (Top 100)

| Users | EER | Users | EER |
|---|---|---|---|
| **1** | 8 | **25** | 18.75 |
| **2** | 12.40 | **26** | 12.5 |
| **3** | 20.39 | **27** | 8.33 |
| **4** | 27.5 | **28** | 10.26 |
| **5** | 15.09 | **29** | 17.94 |
| **6** | 25 | **30** | 16 |
| **7** | 14.71 | **31** | 13.88 |
| **8** | 13.3 | **32** | 4.94 |
| **9** | 16.69 | **33** | 27.77 |
| **10** | 6.45 | **34** | 22.5 |
| **11** | 4.54 | **35** | 3.16 |
| **12** | 9.52 | **36** | 15.90 |
| **13** | 22.7 | **37** | 11.11 |
| **14** | 8.39 | **40** | 20.86 |
| **15** | 19.97 | **41** | 12.40 |
| **16** | 7.73 | **42** | 10.35 |
| **17** | 17.22 | **43** | 17.06 |
| **18** | 13.88 | **44** | 12.5 |
| **19** | 4.05 | **45** | 8.33 |
| **20** | 8.39 | **46** | 17.15 |
| **21** | 8.333 | **47** | 15.98 |
| **22** | 0 | **49** | 13.5 |
| **23** | 10 | **50** | 5.90 |
| **24** | 4.54 | | |

| | |
|---|---|
| | Highest User EER |
| | Lowest User EER |

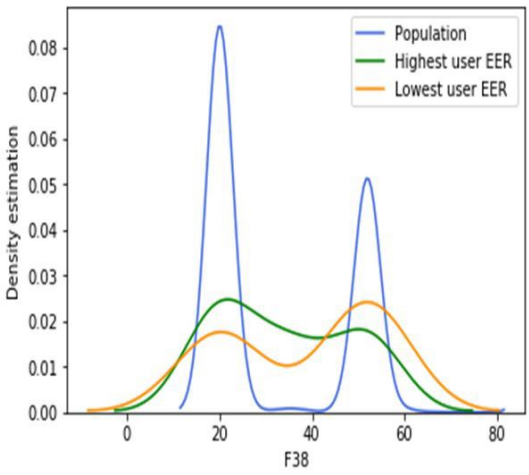
In order to investigate what commonalities and differences exist within the feature set for the population for the Email platform, the top most important features were explored to investigate how these features contribute towards discriminating between the population and thus improve performance.

### ➢ Examining the First Top Thirty Important Features

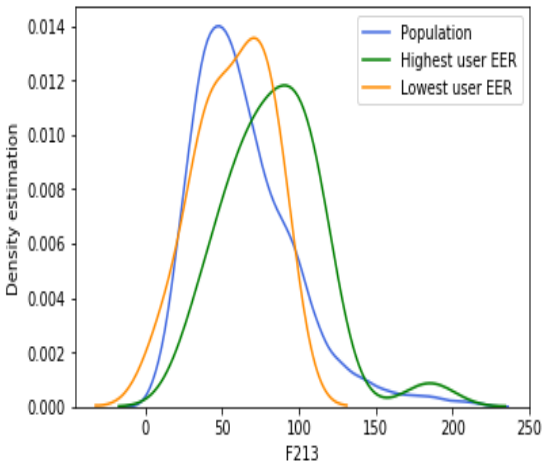
#### • Lexical Features

Further investigations into the Email platform for other lexical features were conducted in order to provide a comprehensive picture and to determine the top lexical features used on this platform when population-based features are considered. It is possible to establish the difference to which input data is similar or dissimilar between authors.

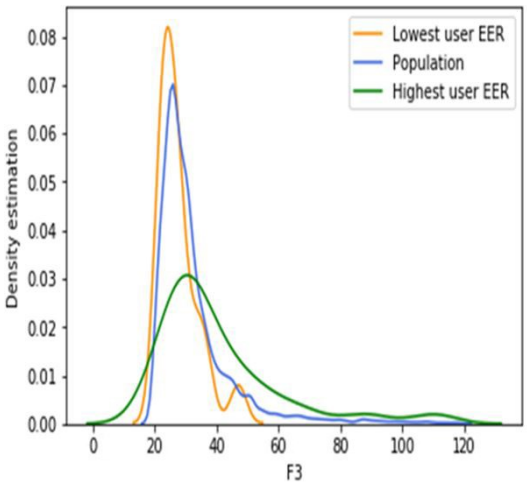
As illustrated previously in Table 4-11, the lowest user error rate was (0%) for *Author 22,* while in contrast, the highest user error rate are was (27.77%) for *Author 33*. The investigation into other important lexical features included: number of special characters, average sentence length in terms of characters, number of uppercase characters, and average sentence length in terms of characters, and average sentence length in terms of words: F38, F213, F3 and F214 respectively. These have been investigated because it has been shown that they are the most important lexical features on the Email platform. By performing density estimation examinations, it is possible to establish the degree to which input data is similar or dissimilar between authors. Figure 4-26 shows the plots of a univariate kernel density estimation to determine the differences in feature distribution for the top most important lexical features between authors; the density estimates show the degree of discrimination for the *population*, the lowest user EER (*Author 22)* and the highest user EER (*Author 33)*.



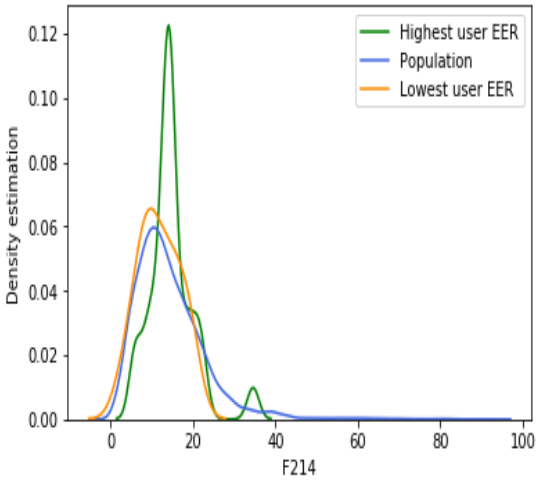

(a) #special characters




(b) average sentence length in terms of characters




(c )# Uppercase characters




(d) Average sentence length in terms of words

Figure 4-26: Density estimation plot for other top lexical features in Emails

The above plots show the similarities and differences in input data between authors, and it can be noticed that the population of authors do not share a similar density distribution, especially with regard to *Author 22* and *Author 33* for (a) *#special character feature*. While it can be noticed that *Author 22* and *Author 33* are similar, nevertheless, they are different, and this simple difference distinguishes them from the *population*, as well as from each other due to their input vectors not being similar or not being within entire boundaries compared to the rest of the population. Thus, this feature is one of the most important lexical

features for discrimination. This led the researcher to explore whether lexical character-based features have some level of discriminative ability for some authors. However, plots (b),(c) and (d) make it slightly difficult to discriminate *Author 22* from the *population,* although they can be discriminated by a chance of 10% because, as the graph clearly shows, the area plot of data coincides between *Author 22* and the *population*. In general, lexical features demonstrate discriminative information and have a level of discriminative ability for the Email platform.

To investigate the way in which authors use short or long words in their Email messages on the Email platform (this feature is explored more in Section 6.5), the distribution of word number for population, user lowest error, and author highest user error were investigated. Figure 4-27 illustrates examples of number of word distribution usage.

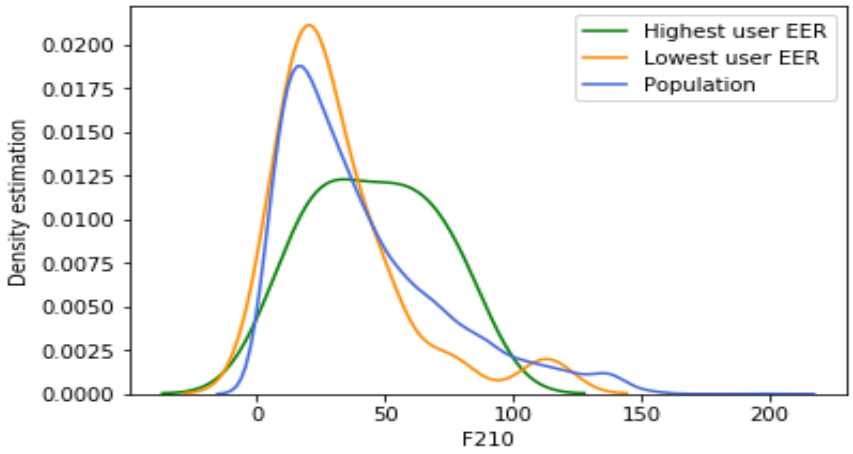



Figure 4-27: Number of words for authors

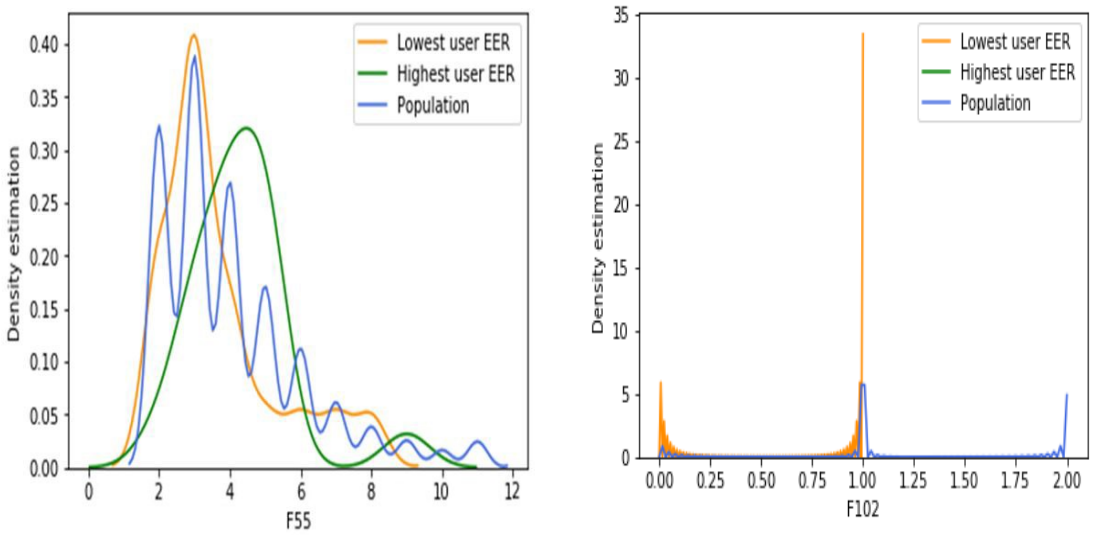
An analysis of the number of words feature indicates that the majority of authors on the Email platform used words that were an average of two to 100 words long in their Emails. However, as can be noted in Figure 4-27, most authors tend to use two to fifty words in their Emails. Also note that *Author 33*, used between 50

and 100 words, which makes them easy to verify and distinguish based on word limit compared to the *population*.

- **Syntactic Features**

  Further investigations into the Email platform for syntactic features were conducted in order to demonstrate a comprehensive picture and to explore the top syntactic features used on this platform. These have been ranked the second most important features on the Email platform, specifically, the feature number of punctuation marks, as shown in Table 4-4. It is possible to establish the difference to which input data is similar or dissimilar between authors, and as illustrated previously, the lowest user error rate was (0) for *Author 22,* while in contrast, the highest user error rate was (27.77) for *Author 33*. The investigation into other top syntactic features included: number of punctuation marks and function words - F55 and F102 respectively, which have been investigated because it was shown that they are the top most important syntactic features from amongst other syntactic features on the Email platform. By performing density estimation examinations, it has been possible to determine the degree to which input data is similar or dissimilar between authors.

Figure 4-28 shows the plots of a univariate kernel density estimation to determine the differences in feature distribution for the these important syntactic distribution features between authors. Density estimating shows the degree of discrimination for *population*, lowest user EER (*Author 22)* and highest user EER (*Author 33)*.




(a) #punctuation (b) Function words.

Figure 4-28 : Density estimation plot for top syntactic features in Email

The above plots show the similarities and differences in input data between the authors. They show that a number of syntactic features have the potential to be used for discriminating users. It can also be noticed that the population of authors share a similar density distribution, *Author 22* and *Author 33* with the *# punctuation feature*. However, it can be noticed that *Author 22* and *Author 33* are not similar, and there was a slight difference with the *population*. This simple difference distinguishes them from population and shows that they are different, as well as distinguishing them from each other due to their input vectors not being similar or not within entire boundaries compared to the *population*. Thus, this feature is one of the most important syntactic features. Although it is slightly difficult to discriminate *Author 22* from the *population,* it is possible to discriminate by a 10% chance, because as the graph clearly shows, the area plot of data coincides between *Author 22* and the *population*. This result provides an important opportunity to advance the understanding of syntactic features in Email, thus, this applies to function words, because, as shown in plot (b), *Author 22* can be discriminated by this feature despite its overlap with the *population*.

Indeed, analysis based on individual punctuation usage indicates that authors used these common punctuation marks such as full stops, exclamation marks, colons, question marks and commas in Email messages, and it is possible to create a profile for punctuation marks on the Email platform because each person has their own style of punctuation.

As expected, most authors are committed to using grammatical rules in Email messages. This is why there was some similarity, but with a slight difference, in this feature among the population. In general, syntactic features demonstrate some discriminative information and have a level of discriminative ability for the Email platform.

To conclude, with respect to the research question concerning the performance of Email recognition across the population, the performance was one of the best after the Text message platform, with an EER of 13.11%. This is probably because it is a private platform and message topics and subjects are similar, with authors using their own writing styles and words.

With respect to investigating the feature vector and how it impacts performance, it can be seen in Table 4-10 that lexical features were repeated more (eighteen times), next were syntactic (nine times), and so these play an important role in discriminating the population and thus improving performance on the Email platform. The next section provides a user-based approach to identify individuals by verifying the authorship of a given text message through exploring users' individual feature sets for different platforms' datasets in detail.

## 4.2 User-Based Approach

This section will explore users' individual feature sets for different platforms' datasets in detail and present the second set of experiments for verifying the

authorship of a given text message. The population-based and user-based approach differ because the population-based approach deals with the classification performance of the population of all users, and uses the ranked feature of the Random Forest algorithm (RF). In population based approaches, the RF algorithm works like a multi-class classification problem to investigate the effect of the top most discriminating features among all the population on the platforms. Whereas a user-based approach involving authorised users can determine the performance of messaging systems with regard to recognition using individual based features, and an RF algorithm can be used to identify only the most relevant features. The Random Forest algorithm deals with this as a two-class classification problem to find the most robust features from a user-base (across authorised users).

The user-based approach involved a repeat of the previous experiment, but based on individual user feature ranking rather than population; therefore, the user-based verification approach has been examined . The main objective is to define the most discriminative user-based features when the user has multiple platforms.

A user profiling technique uses an individual feature set for authors on different platforms, and it is based on an individual feature profiling template. It utilises user features profiling, which is the most distinguished feature for individual use in different text message samples, based on individual linguistic profiles, to create an individual user profile. For example, social messaging platforms provide rich information about individuals, and can be one such source for extracting background information about that individual, for example Facebook (Dewan et al., 2014; Korayem et al., 2013).

A set of experiments were conducted with different settings to investigate the effectiveness of selecting different sets of features for verifying a given user's sample. This includes selecting a varying number of features, ranging from 10 to 275 linguistic features, as an input vector for a classification algorithm. The feature tested 10, 20, 30, 50, 100 and 275 features chosen for this experiment based on the best outcomes from the first experiment; also, the data splitting between training the classifier and testing performance was set to 70/30; 60/40; 40/60; 50/50; 20/80 and 10/90 respectively. Table 4-13 illustrates the overall performance of all authors across the four messaging systems for the four sampling methods. The classifiers chosen for this experiment were based on the best outcomes from the previous experiment: GB, RF and SVM. This is the first study that has attempted to solve the user features across modern platforms together for the author verification problem by exploiting the most discriminating features in this way (to the best of the author's knowledge).

The historical corpora collected are from the four modern corpora of messaging systems that were examined, which are Twitter, Facebook, Email and Text message. The collection and pre-processing methodology of these platforms' corpora are described in detail in Chapter Four. This section is divided into the following: experimental methodology; experimental results analysis, which includes investigating the performance of users on different messaging systems; investigating the performance of platforms with feature vector composition; discussion and, finally, the conclusion.

### 4.2.1 Experimental Methodology

In a user-based approach, the process used to extract features on each user in order to prioritise them in terms of discriminative information, prior to being applied to a standard supervised training methodology has been described

previously in section 4.6 Selecting Discriminative Features. Therefore, to identify only the most relevant user features, the RF deals with this as a two-class classification problem.

Secondly, in terms of user data modelling, as mentioned in the previous chapter, three different classification algorithms were used to find the optimum algorithm for verifying message authorship. This included: Support Vector Machine (SVM), Random Forest (RF) and Gradient Boosting (GB). Each classifier was tested using a different set of features and train/test split ratios. In order to achieve the desired goal and to understand the most important characteristics between the platforms of the user-base, the total number of experiments that have been applied are as follows:

For each platform, each classifier was tested using 36 different sets of configurations (six features tested for each six train/test ratio). In order to investigate the effectiveness of features, as illustrated in Research Methodology Chapter, each of these was repeated according to the number of users in the dataset by using a one-vs-all approach. This resulted in a total number of tests, as shown in Table 4-12 below.

Table 4-12: Total number of tests for all datasets

| Platform | #Users | #Classifiers | Configuration | Total experiments |
|---|---|---|---|---|
| **Twitter** | 41 | 3 | 36 | 4,428 |
| **Text Message** | 26 | 3 | 36 | 2,808 |
| **Facebook** | 46 | 3 | 36 | 4,968 |
| **Email** | 47 | 3 | 36 | 5,076 |
| | | | ***Total** | 17,280 |

In this research, performance was measured based on EER. The experiments were based on 275 features sets, and five different types of linguistic categorised feature sets, which have been categorised as lexical (character and word-based)

features, syntactic features, structural features, short message features and emotional features. The type of feature sets and the selecting of the feature sets is described in detail in Chapter Four.

Thirdly, the under-sampling technique has been used because a common problem that often occurs in authorship verification cases, is the lack of text samples to be used for training, as only limited text samples seem to be available for some authors. This was illustrated in detail in Chapter Five (Section 5.2.2).

### 4.2.2 Experimental Results

In order to select the most desirable user features, these investigations describe changing features with classifiers and settings. Initially, the datasets/features were split into a ratio of 70/30 for training and testing purposes, as illustrated in Table 4-13, and numerous tests were performed (Test-1 to Test-6). The SVM, GB and RF classifiers were applied to train the model on all four messaging platforms. Table 4-13 illustrates the overall performance of all users across the selected features for the six train/test ratio methods tested.

Table 4-13: User-based experimental results (one vs.all approach)

| Test ID | Train/ Test ratio | Feature tested | Performance EER (%) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Twitter | | | SMS | | | FB | | | Email | | |
| | | | SVM | GB | RF | SVM | GB | RF | SVM | GB | RF | SVM | GB | RF |
| Test 1 | 70/30 | Top 10 | 23.78 | 22.02 | 24.22 | 14.38 | 9.04 | 9.69 | 24.59 | 25.1 | 26.31 | 18.04 | 12.95 | 14.41 |
| Test 2 | | Top 20 | 23.4 | 21.16 | 24.01 | 14.23 | 8.17 | 10.37 | 24.95 | **23.78** | 28.33 | 18.75 | **12.03** | 14.46 |
| Test 3 | | Top 30 | 23.12 | 20.53 | 24.95 | 15.5 | 8.18 | 10.46 | 26.81 | 24.08 | 27.69 | 19.12 | 12.16 | 15.06 |
| Test 4 | | Top 50 | 24.39 | 20.42 | 25.74 | 15.72 | 7.99 | 10.54 | 26.14 | 24.64 | 31.45 | 24.15 | 12.33 | 17.66 |
| Test 5 | | Top 100 | 26.95 | 20.45 | 26.48 | 17.27 | **7.97** | 13.23 | 33.27 | 25.13 | 32.13 | 26.41 | 13.05 | 21.03 |
| Test 6 | | All | 31.41 | **20.28** | 29.66 | 21.07 | 8.07 | 16 | 37.92 | 25.09 | 33.58 | 32.85 | 13.87 | 20.36 |
| Test 7 | 60/40 | Top 10 | 23.85 | 22.16 | 23.77 | 15.19 | 9.02 | 11.07 | **25.22** | 26.33 | 26.76 | 18.27 | 13.5 | 14.06 |
| Test 8 | | Top 20 | 23.74 | 21.46 | 25.76 | 15.52 | 8.44 | 12.26 | 25.65 | 27.04 | 28.11 | 18.83 | 14.1 | 15.41 |
| Test 9 | | Top 30 | 24.29 | 21.16 | 24.14 | 16.34 | 8.27 | 10.3 | 27.58 | 26.75 | 30.4 | 21.24 | 13.75 | 17.04 |
| Test 10 | | Top 50 | 25.25 | 20.79 | 26.89 | 17.48 | 8.16 | 11.96 | 28.99 | 27.65 | 31.7 | 21.9 | 13.49 | 17.95 |
| Test 11 | | Top 100 | 28.3 | **20.38** | 27.33 | 20.57 | **8.11** | 13.15 | 33.77 | 26.71 | 32.89 | 29.33 | 13.46 | 21.18 |
| Test 12 | | All | 31.33 | 20.47 | 28.81 | 23.17 | 8.29 | 16.06 | 40.72 | 26.62 | 34.27 | 32.98 | **13.34** | 24.5 |
| Test 13 | 40/60 | Top 10 | 23.79 | 23.11 | 24.91 | 15.02 | 9.1 | 11.7 | 30.82 | **28.97** | 29.81 | 23.36 | 17.16 | **16.16** |
| Test 14 | | Top 20 | 25.68 | 22.41 | 28.18 | 15.56 | 8.75 | 13.31 | 31.45 | 29.28 | 31.83 | 24.78 | 16.53 | 17.46 |
| Test 15 | | Top 30 | 25.39 | 22.22 | 26.18 | 16.29 | 8.46 | 12.41 | 34.18 | 30.13 | 33.71 | 24.11 | 16.33 | 19.71 |
| Test 16 | | Top 50 | 28.9 | 21.88 | 25.69 | 17.52 | 8.32 | 13.2 | 34.18 | 29.48 | 33.99 | 27.47 | 17.68 | 21.61 |
| Test 17 | | Top 100 | 32.29 | **21.75** | 30.31 | 20.94 | **8.3** | 14.01 | 36.73 | 29.82 | 35.75 | 30.68 | 18.04 | 23.51 |
| Test 18 | | All | 34.86 | 21.86 | 32.19 | 22.97 | 8.36 | 17.64 | 44.29 | 29.82 | 36.01 | 37.75 | 18.45 | 26.69 |
| Test 19 | 50/50 | Top 10 | 23.92 | 22.82 | 23.97 | 15.7 | 8.86 | 12.34 | **25.75** | 28.59 | 29.48 | 17.93 | 14.63 | 15.59 |
| Test 20 | | Top 20 | 23.58 | 21.49 | 26.28 | 15.4 | 8.45 | 12.8 | 26.61 | 27.03 | 32.58 | 20.71 | 14.33 | 18.01 |
| Test 21 | | Top 30 | 25.22 | 21.17 | 25.84 | 16.11 | 8.27 | 12.78 | 26.7 | 27.43 | 32.25 | 22.32 | 13.96 | 18.2 |
| Test 22 | | Top 50 | 26.96 | 21.18 | 27.35 | 17.45 | **8.11** | 13.06 | 30.53 | 27.97 | 32.74 | 26.22 | 14.74 | 17.93 |
| Test 23 | | Top 100 | 29.37 | **20.85** | 28.91 | 20.14 | 8.13 | 14.79 | 37.83 | 28.05 | 34.23 | 31.22 | 14.16 | 22.95 |
| Test 24 | | All | 33.82 | 21.09 | 30.13 | 22.74 | 8.16 | 16.67 | 42.74 | 27.87 | 35.31 | 37.16 | 14.27 | 24.22 |
| Test 25 | 20/80 | Top 10 | 35.66 | 27.51 | 29.23 | 21.76 | 11.03 | 12.74 | 37.38 | 34.58 | 35.11 | 32.54 | **22.23** | 23.67 |
| Test 26 | | Top 20 | 35.19 | **27.13** | 30.46 | 23.4 | **10.88** | 15.96 | 39.87 | **33.33** | 36.75 | 35.37 | 23.82 | 25.59 |
| Test 27 | | Top 30 | 35.45 | 27.97 | 32.3 | 24.23 | 11.71 | 13.07 | 37 | 34 | 38.47 | 36.12 | 24.03 | 25.5 |
| Test 28 | | Top 50 | 34.54 | 27.78 | 32.01 | 24.78 | 11.17 | 15.48 | 38.92 | 35.27 | 39.47 | 36.28 | 25 | 26.33 |
| Test 29 | | Top 100 | 35.17 | 28.33 | 35.25 | 26.51 | 11.38 | 20.28 | 45.47 | 34.96 | 39.49 | 42.56 | 25.93 | 29.67 |
| Test 30 | | All | 37.78 | 28.59 | 34.69 | 29.02 | 11.54 | 21.74 | 50.44 | 35.12 | 41.13 | 48.45 | 26.25 | 32.41 |
| Test 31 | 10/90 | Top 10 | 39.44 | **31.43** | 32.84 | 25.31 | 12.03 | 16.01 | 48.58 | **38.86** | 39.4 | 47.79 | **27.19** | 27.19 |
| Test 32 | | Top 20 | 38.68 | 33.6 | 35.01 | 25.5 | **11.74** | 19.67 | 51.26 | 38.89 | 39.58 | 49.61 | 28.36 | 31.13 |
| Test 33 | | Top 30 | 39 | 33.78 | 35.88 | 26.64 | 12.03 | 17.6 | 50.58 | 39.03 | 41.31 | 47.77 | 28.62 | 32.28 |
| Test 34 | | Top 50 | 39.35 | 34.48 | 37.11 | 27.24 | 11.77 | 18.11 | 49.86 | 40.08 | 41.64 | 48.08 | 29.35 | 34.21 |
| Test 35 | | Top 100 | 43.33 | 34.21 | 37.34 | 29.57 | 12.37 | 21.36 | 53.51 | 38.99 | 42.86 | 51.57 | 30.09 | 35.08 |
| Test 36 | | All | 46.96 | 34.48 | 39.51 | 32 | 14.43 | 23.26 | 55.89 | 39.73 | 43.8 | 53 | 30.66 | 37.06 |

It can be observed that during the practical experiments, the GB classifier performed better on Twitter, Text message, Email and Facebook, although SVM may be relatively better regarding the performance of Facebook. Therefore, from the practical experiments described, one of the main findings noted was that by increasing the stylometric feature set, the EER increased, as shown in Table 4-13 below. It can also be noticed that during user experiments, these preliminary results seem to indicate that there is a real subset of common features that can be shared between the Email and Text message platforms, as the EER of email resulted 12.3%, and the EER of Text message was 7.79%. The worst results out of the platforms were for Facebook and Twitter, as shown in Table 4-13. This might be because the individual's text messages are often sent to the public, and individuals commonly tweet and post, which may be why the classifiers struggled to perform well on these messaging systems through the different features tested, and it indicates that there was more variation. There is another reason why these platforms showed poor in performance for individuals, which is because copy and pasted text messages between users on these platforms are significant (Farahbakhsh et al., 2016; Ottoni et al.,2014). While in contrast, Text message and Email mostly contain private messages - people send text messages to known persons and these usually cannot be copy and pasted from others.

The following consists of a set of investigations that were conducted to address the core research questions related to the first part of the research equation on understanding user performance and messaging systems recognition using multiple text message samples, as well as how this performance compares across platforms. Finally, a preliminary discussion of the possible common stylometric features between platforms can be found at the end of this chapter.

In investigating the performance of users on messaging systems, several experiments were conducted to explore the research question proposed by examining the reliability of user recognition when dealing with multiple messaging systems for the user base. Then the results from all platforms were analysed in the next trial as follows:

- The effectiveness of number of features for classification performance was explored (i.e. top 10 features, top 20 features to top 275 features).
- The effectiveness of ratio of train/test changes for system performance (i.e.70/30 train/test) was tested.
- The impact of user performance on messaging systems recognition using different text message samples, and how this performance compares across platforms, were examined.
- The commonalities and differences that exist within the feature set for user-based similarities were analysed, leading to a preliminary discussion of the possible common stylometric features between platforms at the end of this chapter.

The purpose of these investigations was to explore if there is any impact from the number of features on performance in the case of there being not enough features to investigate the suspects. Therefore, from these results it can be concluded that the best ratio for gaining optimum results is from setting the train/test ratio to 70/30, as shown in Table 4-13. It has also been noted that for Twitter and Text message, the best performance was reached by using all the features, or at least the top 100 features, due to the nature of these platforms, since they have small capacity and writing ability is limited and more writing is needed to verify users. While for Facebook and Email, the top 20 features is significantly sufficient to achieve better performance.

With respect to the second investigation into the impact of user performance on messaging systems recognition using multiple text message samples, and how this performance compares across platforms, the experiments show that Text message and Email achieved good performance at of 7.97% and 12.03% respectively. While Twitter and Facebook messages achieved poor performances, with an EER of 20.28% and 23.78% respectively. An analysis of the dataset in terms of size and the composition of individual users on the Text message and Email platforms shows that the individual author is likely to use the same words, characters and writing style (i.e., authors use the same writing style and vocabulary when writing on these platforms), and there is a clear indication that the writing style used between these two platforms is likely to be similar, as was explored previously in the population based section. For example, the writing is characterised by certain features or private vocabulary or personal word reference for platforms by the user, and texts directed to specific known people and size are less likely to be a determining factor in the composition of the message itself. Thus, the results show a significant difference in performance depending on the platform being analysed.

On the other hand, and in terms of feature testing, subsets of stylometric features are more reliable in determining authorship using a few features, as they have performance with a few features, such as 7.97%, 12.03%, and 23.78% for Text message, Email and Facebook platforms respectively. In terms of feature vectors, Facebook and Email are more verifiable and can be verified with only a few features (the top 20 features) since the user often writes a longer message on these platforms. Often, a lot of writing gives a wider indication of user recognition; whereas linguistic tendencies are often determined and vice versa with Twitter and Text message, and these require more features to be verified for short texts.

This is the first study of its kind, as there has not been any previous research that has examined stylometric features and their relative performance across four modern platforms together (to the best of our knowledge).

In general, after the performance showed successful results, determining the best features for authors on each platform was attempted based on the best performances and feature tests, as shown in Table 4-13 (e.g. the best performance for Text message was an EER of 7.97% with the top 100 features). Table 4-14 below shows the authors' performances on the platforms Twitter, Text message, Facebook and Email. To illustrate the performance of each user on each platform in a simplified way, the performances are coded in colour, where the red represents a high performance for the user, while green represents lower performance; while white represents the average performance or no performance of the user. The subset features for authors on different platforms were determined, including features that are shared by multi-platforms, and this is discussed in the next section.

Table 4-14 : Users' performance with different platforms

| EER(%) performance on platforms | | | | |
|---|---|---|---|---|
| User | Twitter | SMS | Facebook | Email |
| 1 | 17.8 | 5.6 | 20.9 | 10 |
| 2 | 0 | 5.4 | 17.7 | 12.4 |
| 3 | 21.4 | 0 | 11.8 | 25.2 |
| 4 | 32.5 | 12.7 | 6.3 | 11.3 |
| 5 | 33.3 | 6.8 | 27.3 | 13.4 |
| 6 | 30.4 | 12.5 | 38.9 | 25 |
| 7 | 31.4 | 11.8 | 20.7 | 17.3 |
| 8 | 17.9 | 12.5 | 20.9 | 16.7 |
| 9 | 26.3 | 16.2 | 26.1 | 16.7 |
| 10 | 23.9 | 12.7 | 39 | 12.9 |
| 11 | 20.6 | 4.9 | 14.9 | 4.5 |
| 12 | 25 | 6.4 | 17.4 | 7.4 |
| 13 | 28.3 | 3.6 | 21.6 | 15 |
| 14 | 26 | 13.2 | 36.1 | 3.8 |
| 15 | 18.2 | 0 | 30.2 | 25.3 |
| 16 | 28.2 | 5.4 | 32.8 | 7.7 |
| 17 | 28.3 | 14.2 | 4.8 | 13.8 |
| 18 | 0 | 7.3 | 16.5 | 9.7 |
| 19 | 25.3 | - | 31.6 | 3 |
| 20 | n/a | 4.5 | 27.6 | 8.4 |
| 21 | 12.1 | - | 16.7 | 0 |
| 22 | 34.2 | - | 26.1 | 0 |
| 23 | 13.2 | - | 22.9 | 17.1 |
| 24 | n/a | 13.7 | 12.5 | 4.5 |
| 25 | n/a | 8.8 | 19.2 | 0 |
| 26 | 24.7 | - | 21.8 | 20.8 |
| 27 | n/a | 0 | 27.6 | 8.3 |
| 28 | 34.2 | - | 19.4 | 15.9 |
| 29 | 10.5 | - | 6.5 | 16.7 |
| 30 | 0 | 3.9 | - | 8 |
| 31 | 17.3 | - | 24.2 | 5.6 |
| 32 | 4.9 | - | 25.6 | 5.8 |
| 33 | 19.8 | - | 33.3 | 11.8 |
| 34 | 22.7 | - | 5 | 15 |
| 35 | 15.3 | - | 30.5 | 3.8 |
| 36 | 32.3 | - | 37.5 | 15.9 |
| 37 | 34.6 | - | 33.3 | 11.8 |
| 38 | - | 3.7 | 8.4 | - |
| 39 | 8.3 | 14.2 | 6.3 | - |
| 40 | - | 7.2 | 36.8 | 25.5 |
| 41 | 32 | - | 35.6 | 16.3 |
| 42 | 12.5 | - | 33.4 | 13.8 |
| 43 | 18.1 | - | 24.2 | 17.1 |
| 44 | 12.5 | - | - | 6.3 |
| 45 | 7.1 | - | - | 8.3 |
| 46 | - | - | 36.4 | 14.3 |
| 47 | - | - | 29.5 | 20.9 |
| 48 | 23.4 | - | 30.5 | - |
| 49 | 7.1 | - | - | 16 |
| 50 | - | n/a | 27.6 | 6.5 |

Table 4-14 shows that some authors perform better on more than one platform, whereas some authors showed poor performance. An overview shows that the worst authors' performances are for Facebook and Twitter, while in contrast, Email and Text message show the best performance. Although some authors have achieved an EER of 0% on the worst platforms of Facebook and Twitter, such as *Author 2* and *Author 18*, the majority of authors showed poor performance. In contrast, the other group shows that some authors achieved good performance on the Text message and Email platforms, although most authors did not achieve 0% EER; however, the majority of authors had better results on Text message and Email compared to Facebook and Twitter, for example *Authors 1,5,6,7,* and *16*. It may be possible to determine that the best performance for others, and closer relative performance to each other, are as follows: 1 – Text message, 2- Email, 3- Twitter, and 4- Facebook. This indicates that there were some common features shared between corpora for some authors.

## 4.3 Discussion

Although the nature of real data for each platform is considered to differ from one another, and the stylometric features vary from person to person, the results achieved are promising. The results reflect a high possibility of deploying the proposed forensic investigations to compare data across platforms to support existing active messaging systems for crime investigation, such as the writing style of suspects. Since there is currently no real-life composite or a multi-platform dataset in the messaging systems field, the comparison with related works is relatively limited. As has been reviewed and discussed extensively previously in the literature review, most of the previous studies have focused on a single platform and have attempted different feature techniques by using various single corpus, and different limited numbers of authors; therefore, it is still not clear how

to identify the most appropriate features, and in the majority of studies, the number of samples is few and they involved different complex techniques compared to the large number of samples for more than one platform and approach in this current study.

The results of the performance in both experiments (as illustrated in Table 4-2, and Table 4-13) for population and user-based features for single platforms are positive, and it is encouraging that there are some strong features, such as lexical features, which may provide common features. They are the most powerful features, and within the category of lexical features, character and word based is the closest feature and is the most effective category for lexical across platforms that could be further investigated concerning the feature commonality of population and individual authors across corpuses. It is difficult to find common features between platforms if there are not more than one platform for which data has been collected and extracted, hence, as mentioned previously, this research has targeted authors who have at least two platforms. Moreover, this research has utilised four historical datasets containing a large number of messaging system samples (4,539 samples for Facebook, 13,616 for Twitter, 6,538 for Email and 106,359 for Text message) across more authors (i.e. 50 users), and it has covered most scenarios to assess the ability to compare across platforms. In addition to collecting different types of historical data, more terms and conditions have been implemented, including authors having to have at least two platforms and no less than 20 text messages on all platforms, without knowing the age of the account or the maximum number of text messages, which is because most users do not remember when they created an account, for example SMS, offering the opportunity to learn the user's linguistic behaviour in a more realistic way, rather than under laboratory conditions or by calculating their range of vocabulary.

Moreover, different linguistic features were extracted from different core modern messaging systems, including comprehensive modern social networks features, for example Facebook and Twitter, which contributed towards the creation of a larger feature vector for the linguistic and forensic domains for each platform in comparison to prior research, and the performance was better, with the best EER of 7.97%, 12.03%, 20.28% and 23.78% for Text message, Email, Twitter, Facebook, respectively, suggesting the potential usefulness of the proposed method. In addition, the three classifiers RF, GB and SVM were utilised and their impact on the system's performance investigated, such as feature tested (i.e., top 10, 20, 30, 50, 100 etc.). This is the first study of its kind, as there has not been any previous research that has examined stylometric features and their relative performance across four modern platforms together (to the best of our knowledge). In summary, these results show that:

From both a population and an individual classifier performance perspective, the experiments show that a user-based feature profiling approach has performed better than a population-based feature profiling approach. This was expected since the input vector contains only strong discriminating features for each individual.

The majority of the most common features on all single platforms for population-based features were lexical, as shown in the population experiments. The exploration of feature vectors has been analysed based on the performance of each single platform (i.e. the top 30 features for the Text message platform with an EER of 7.97%, with GB as the best classifier), and visualised by performing density estimation examinations to explore the discriminative power of this top feature and how it impacts on performance, as well as the degree to which input data is similar or dissimilar between populations. This includes a comprehensive

survey conducted linguistically within platforms for which these subsets of stylometric features would be more reliable in determining authorship among the population. Also, the top categories for each single platform for population base (i.e. Number #special character F31 in Twitter) have been analysed.

The results show a significant difference in performance depending on the platform utilised. Lexical features show a positive impact on most platforms (Text message, Email, Facebook and Twitter, respectively), as shown for each single platform in Table 4-4, Table 4-2,Table 4-8, and Table 4-10 for Twitter, Text message, Facebook and Email respectively.

A user-based technique has played a major role, and has contributed towards determining that each individual has their own unique writing style and linguistic behaviour features across platforms; for example, the feature *number of alphabet a-z* for *User 1*, as per the example shown in Figure 4-1. This can be extended to other common features with other different messaging systems. When reviewing the performance of users across messaging systems, it has been found that authors may use common feature sets across platforms, as shown in Table 4-14. This could help the classifier to identify the user more easily because they appeared strong when the strongest (ranked) features were captured across platforms (e.g. the features of *User 1* appeared on Facebook and Twitter), and they differ among authors. A user that performs well on one platform does not necessarily have any direct correlation to a user that performs well on other platforms. This is the first study that has attempted to solve the author verification problem across modern individual corpora together by exploiting the most discriminating features of authors using multi-class classification.

An analysis of the dataset shows that in terms of size and composition, Text message and Email repositories represent either end of the spectrum, suggesting volume is less likely to be a determining factor over the composition of the message itself. The results show a significant difference in performance depending on the platform. With respect to the research question: Understanding the performance of messaging systems recognition for population and user-based features), the Text message platform achieved the best performance compared to the other platforms in all scenarios at 7.79%. Followed by the Email platform at 12.03%, then Twitter at 20.28%, and finally Facebook at 23.78%. As shown in Table 4-13, to determine the best classifier intersecting with the most distinctive stylometric features for each single corpus, 36 tests were conducted.

- With respect to the research question regarding exploring the feature vector and how it impacts performance: What commonalities and differences exist within the feature set for individuals? In addition, what commonalities and differences exist within the feature set across the platforms? The exploration of the feature vector has been analysed for individuals across four platforms, and examples are given in Figure 4-1, Figure 4-2 , Figure 4-3, Figure 4-4, Figure 4-5 and Figure 4-6 (section4.1.3), and the strength of these features for individuals is discussed in the next chapter.
- With respect to the research question concerning what commonalities and differences exist within the feature set for population versus user base, it has been shown that the user-based feature profiling approach performed better than other profiling techniques, since the input vector contains only strong discriminative features for each individual author. While the

population treats all strong features equally and classifies the strength of features based on their distribution among all users.

- A number of experiments were conducted to investigate the GB, SVM and RF classifiers for both the population and user base. It can be seen in Table 4-2 and Table 4-13 that better results were achieved by the GB, which outperformed most prior studies and never has been used across modern platforms together. Broadly speaking, the SVM classifier was identified, as it has achieved good performance with various domains of author verification on single platforms and unclear mechanisms, and with limited datasets, as described in the Literature Review chapter (Zheng et al., 2006; Green et al., 2013; Li et al., 2014; Allison et al. 2008). The reason for considering SVM classifiers out of other classifiers in most prior art maybe because: SVM has strongly contributed and can play an active role, especially with a small data size. This indicates that it has outperformed other classifiers, which is why it may be the best classifier when a small volume of data with a limited number of users is used, especially for a single corpus.

### 4.3.1 Comparison with the Prior Art

As identified in Chapter Three, most of the previous studies have focused only on one platform's potential, and a lot of work on author verification from text on different platforms has been undertaken. However, no research has been found that seeks to employ multi-modern four platform and multi-features for identifying features that can lead to author verification (such as multi-features on variety platforms) to advance the state of knowledge and enable a better decision-making process. As a simple comparison, none of the previous systems has attempted to cover a wide variety of real-world datasets for various messaging

systems, that is, studying the potential features of multiple platforms - Twitter, Facebook, Text message and Email -under realistic circumstances and with real text sample combinations. Accordingly, there is a need to propose and to figure out how the author can be verified forensically if he or she uses more than one platform, as a suspect may use one platform in a kind and positive way with people, but use another one to spread serious threats and hate that impacts on people online and society in general.

The dataset is an essential part of the verification process, however, some corpora are publicly available, such as Facebook and Twitter, allowing the offender to send text messages to the public easily because it is simple and without restriction or control. On the other hand, there may be no Text message or Email messages available to assess due to their high level of privacy. In the current study, the historical dataset contains over 131,054 samples collected from across four corpora from 50 subjects. Also, each participant had to have at least two data platforms available.

Therefore, a compression of the study into a single platform with their individual platforms (one platform to one platform) was reasonable. Furthermore, no previous research has explored integrating the features of multi-platforms to verify authors through common feature analysis. Also, no research has examined a real dataset in this way (to the best of my knowledge). Although there are lack of studies that have used real samples for platforms, attempts have been made to try to verify authors using different techniques and improve the result using different methods. Hence, to compare the results of previous studies with this study for one platform against another, a comparison (one platform against another) has been made and the results are summarised: In terms of Text

message investigation, as presented in Chapter 3, there have been a lack of studies on the Text message platform.

The most seminal research study in this field was conducted by Saevanee and Clarke (2011). Their research achieved an EER of 24%. Their findings are based on 30 participants, with minimum of 15 samples per user, but maximum samples were not mentioned and neural network RBF was a classification rate. While their results are desirable, the situation could be different if the technique was applied to the top discriminatory features or most effective stylometric features, as that could increase the level performance and this is non-existent compared to this study for their single corpus.

In terms of Twitter, as stated in Chapter 3, there are a lack of studies on this platform. The most seminal work in this field was conducted by Brocardo et al (2017). Their research study achieved an EER of 16.73% for 10 authors and 100 samples per author. Lexical, syntactic, and application-specific features were utilised as feature set.Their technique relied on the n-gram technique to measure the degree of similarity between a block of characters and the profile of an author. Although the number of features, number of authors and samples are small compared to this research, their results seem to be slightly better, as while they included a small number of authors which is10 authors and few of samples, the slight increase in EER may have been due to the use of the n-gram technique with a low number of authors and few samples. The n-gram technique may not be suitable for use with large numbers of samples since it was designed to deal with a small dataset. The mechanism of n-gram involves calculating the number of serial and sequential words and letters for a specific author, but it does not represent the nature of the text for that author since it only performs a calculation without knowing and understanding the nature of the text and the features used

by authors. Knowing what drives decisions on features (i.e. the features on which the investigation relies) is an important element in some messaging recognition applications, such as courts and crime-related research. Secondly, n-gram features can be noisy since tweets are non-structured. Thirdly, the suspect can simply change his/her writing strategy, and this change will undoubtedly affect the calculation process used in this technique, meaning that it will not be accurate since it does not deal with the text features of that author.

In terms of Facebook investigations, the most prominent previous research study of Facebook platform was by Li et al (2016). They used SVM Light as the classifier type with 233 features, and their accuracy was 79.8%, with an EER of approximately of 20%. While the performance in this study was slightly less, an EER of 23.78% due to the number of participants in their study less, 30 participants, and this contributes to improve the performance if the number of participants is less.

In terms of Email investigation, the most prominent previous research study of Email platform verification is by Iqbal et al, (2010), which yielded an EER ranging from 17.1% to 22.4%. One idea in their study was to cluster anonymous Emails by using stylometric features and extracting the write print to verify the author. However, their technique is based on clustering and mining the writing styles from a collection of Emails written by multiple anonymous authors, and attempting to group Emails written by the same author.

There are many aspects that may have had an impact on their EER, and caused the EER of this research to be better than the EER found in their studies, for example: firstly, they used clustering and mining of writing styles, which means they did not deal with the most effective and discriminating features for the Email

platform and for every participant in order to limit discriminating features; however, there is no need to do so, since the selection of robust features would help to determine the strengths and weaknesses of that platform in order to make a strong verification and acceptable user EER. Secondly, the classifier of SVM yields a low EER compared to the GB classifier. Finally, there is no study (to my best of knowledge) to date that has investigated the stylometric features of these electronic messaging platforms (Facebook, Twitter, Text message and Email) joined with each other for purposes of comparison.

## 4.4 Conclusion

This research study has sought to investigate the relative performance of linguistic feature recognition across a wide range of independent modern messaging systems for the population and individual users. Based on 50 participants, the investigation has provided significant evidence to suggest that the ability to compare across platforms using common linguistic features is a reliable means of cross-platform assessment for verifying the population and users of multiple platforms. There will be challenges across these platforms, as there is no clear data on the best platform or what features are best on all of them. From the exploration of data feature vectors, they work well for some users on some platforms, but there is relatively little information to suggest that good performance on one platform will be good on another.

On average, for a population-based approach, the best performances of platforms for the experimental results achieved was for Text messages, with an EER of 7.97%; followed by Email with an EER of 13.11%; then, Twitter tweets, with an EER of 20.16%. Finally, the worst performance from all four platforms and categories was the Facebook platform with an EER of 25%. This shows the usefulness of single-domain platforms where the use of linguistics is likely be

similar, for example Text messages and Emails, which have more in common, specifically lexical features.

For the user-based approach, there is very little evidence to suggest a strong correlation of stylometry between platforms, meaning that users communicate quite differently with different sets of stylometry on individual platforms. However, it has been found in this current research that the best experimental results achieved were for Text message, with an EER of 7.97%, and three authors experienced EERs equal to or less than 0.2%; followed by Email with an EER of 12.03%; then, Twitter tweets, with an EER of 20.28%. Finally, the worst performance from all four platforms and categories was the Facebook platform with an EER of 23.7%. The best ratio for gaining optimum results is when setting the train/test ratio to 70/30 compared to all other tested six settings, and the best classifier was the GB classifier compared to the other three classifiers tested for both population and user base on modern platforms jointly.

This evidence suggests that linguistic features on individual platforms such as text messages have features in common with other platforms such as Email, and lexical features play a crucial role in the similarities between users' modern platforms.

Many stylometric features have been suggested in previous studies for only one single platform for authorship verification, for instance, the choice of lexical and syntactic, structure features and so on. However, it is not clear which of the stylometric features would be robust and trusted enough in the case of combining more than one platform together.

Many studies have also tried to use text messages that are not real or arranged with participants, which caused them to have to write specific text messages

artificially, rather than being real; whereas in this study, the messages are real and were collected from Plymouth University students without requesting them to write messages before they came, because the purpose of the study is to seek the real linguistic features of each platform without prior agreement on the quality of the messages.

The analysis above provides strong evidence and indicates that a number of features could be very useful for verifying authors from a population and user base. For example, lexical features show a very strong discriminative element for some authors in the population. As demonstrated by the population-based features, each author shows a degree of uniqueness when selecting the top discriminating feature for the population that underlies the behavioural characteristics of the language on each platform. Population features only determine the robust features for the level of population or platform, while they do not determine the robust features between authors themselves, because every author has their own unique linguistic features. By using all discriminating linguistic features for everyone (e.g. a user-based verification approach), the input vector contains more discriminatory information to differentiate a user and result in good classification performance. Therefore, a user-based feature verification approach has also been considered to address this problem. It has been demonstrated that the classification performance will be improved by using a user-based feature approach because only the selected features are used as part of the input vector, but these are dependent on individual analysis rather than the population.

This chapter has also discussed a stylometric features technique, for which features can be used for cross-platform authorship. Moreover, it has been shown

that lexical features have the ability to be investigated across messaging system platforms, as well as it working with the Email and Text message platforms.

Importantly, no prior study or research has collected data from the same users on up to four platforms. Therefore, this research is the first to compare directly across four platforms, and it has also explored feature vectors and looked at population and user-based data, along with further investigations to give the ability to compare. On the other hand, from a biometric perspective, the enrolment process in feature analysis requires the existence of an enrolment sample, which is used to compute the behavioural profile of the user. This sample should contain all possible key combinations in order to effectively recognise the user based on an expected or unexpected set of author inputs. The ability to verify the user does not only have application in the digital forensic dominion, but could also be used as a biometric system modality for use in transparent authentication.

Therefore, it is important to investigate authorship verification in a platform-dependent manner, and to compare the relative performance of author verification across multi-short messaging platforms, including assessing how well author verification performs across platforms platforms, and exploring feature vector composition, as well as the impact of classification on performance. The next chapter will provide more depth by analysing the stylometric feature vectors of different modern platforms, both single-domain and across-domain. It is divided into six sections: the introduction; feature vector composition analysis, which contains cross-platform authorship among population and user-based; unified feature profile; feature vector portability; message lengh performance, and finally, the conclusion.

# 5. Chapter Six: Platform-Dependent Author Verification

## 5.1 Introduction

As has been shown in the previous chapter, this research has involved collecting data from up to four platforms from the same users to understand feature analysis across modern platforms, and to provide the opportunity to compare directly across those platforms. After investigating the performance and feature vectors across platforms, and having looked at the population and user base for independent author verification in the previous chapter, this chapter will explore in detail the second and the third proposed research questions on how well authorship verification will operate across platforms in both single-domain and cross-domain datasets, as well as message length performance. In addition, it will address whether there are any common stylometric features between platforms for both the population-base and user-base in single-domain dataset verification approaches and cross-domain datasets. From a biometric perspective, this chapter can be defined as presenting descriptive statistics that allow the nature of the data to be explored in order to find out what commonalities in data exist across platforms for platform dependent author verification.

It should be noted that platform independent refers to looking at feature vectors from a population perspective versus an individual user perspective independently (separately or individually), having looked at the population and user base for independent author verification for Platform-independent, and Platform-dependent, Chapter Six explores how well authorship verification can operate across platforms.

The experiments consist of three methods: the first method is feature analysis involving verifying the authorship of different text samples in single-domain

datasets for both the population and user-base; the main aim of the first method is to explore common features across platforms. The second method unifies all features from different messaging platforms, for example unifying the most discriminating features for Twitter, Text message, Facebook and Email. This means unifying the author's top discriminating features from across platforms (unified linguistic features of an author for multi-platform verification). The main aim of the second method is:

- To discover whether there are any common features, if they exist, when platforms are unified.
- To help in finding a systematic and forensically automated method to be used under one umbrella mechanism with he potential to assist linguistic experts to create profiles of authors flexibly and reliably across platforms.
- To discover the possibility of assisting forensic experts to identify the movements of features that may contribute towards tracking the features of an author's profile across platforms (identifying proper features across platforms should lead to identifying the author) and to support intelligence applications to analyse aggressive and threatening messages.
- To find different ways of unifying subsets of common user features across platforms (explore how a profile can be unified across platforms).
- Finally, to explore unifying author features across platforms in order to understand what it is the most powerful feature, if any, and if it exists within linguistics cross-domain.

For the third approach, portable features across platforms (portability approach) have been used to verify the authorship of different cross-domain dataset samples. The main aim is:

- To explore the portability of author features across platforms in order to understand what feature, if any, is the most powerful, and if one exists within linguistics cross-domain; along with assessing the potential for future research in this area. Most previous techniques concerning authorship verification have assumed that the training and test data are drawn from the same distribution, but this novel research is different as it uses real scenarios. In addition, it uses cross-domain settings, that is, testing Facebook posts versus testing Twitter tweets (portability linguistic user features verification).

Furthermore, by exploring common features across platforms (feature analysis), and unifying the top features and portability of the discriminative features of an author, it may be possible to find the common features across platforms. As shown previously, Figure 0-2 reflects the feature spaces of the second and the third approaches to highlight which feature was being used.

The datasets collected from four different messaging systems have been examined in Chapter Five. In addition, the methodology used for the data collection and pre-processing of these platform datasets is described in detail in Chapter Four, and the population and user-based verification approaches were examined in the previous chapter and provided the opportunity to compare directly across those platforms. The feature analysis-based user verification approach across-platforms for population and user-base; the unified feature-based user verification approach, and the holistic portability user feature across platforms will be investigated in this chapter.

## 5.2 Feature Vector Analysis

The previous chapter has shown that each platform has a degree of uniqueness (top stylometric) and a feature that underlies the behavioural characteristics of the platform's language. For example, lexical and syntactic features were noticed more on Twitter when the top thirty features were captured, whilst the top features on the Text message platform showed that the most inflectional features were lexical, syntactic structure and emotional. The analysis included the features of some of the authors for different messaging systems (Twitter, Text message, Facebook and Email). The following sections present an analysis of the experiments conducted to address the core research questions that are related to common features among the population (i.e. features across and between platforms) and the user-base. Therefore, in order to explore to what extent stylometric features are common between multiple messaging systems, and to obtain sufficient information to create the reference template, the common stylometric features for the top features for each platform have been explored. After verifying the most influential features across platforms by using the Random Forest algorithm, this section is divided into the following subsections: population-based analysis (Common Feature Vectors among the Population), and user-based analysis (Common Feature Vectors that are User-Based).

### 5.2.1 Population-Based Analysis (Common Feature Vectors among the Population)

An analysis of the top features for population for each platform has been conducted in order to explore the most common features among them. The top features were captured for authors after ranking them using the RF algorithm. The top 10 features, including its category; the top 20 stylometric features, and finally the top 30 stylometric features for each population platform have been analysed.

The reason for selecting the top-most features is because it should show that there are some common features across the first ten features, and the first twenty, and to ensure the first thirty features have also been explored. Also, the least tested features that have performed well on one of the platforms for population base is the top 30 for the Twitter platform, as shown inTable 4-2. Three approaches have been used, which are:

- Feature analysis of the top 10 most stylometric feature for each platform
- Feature analysis of the top 20 stylometric features for each platform
- Feature analysis of the top 30 stylometric features for each platform

The top stylometric features for each platform have been investigated in order to see whether it is possible to find common features among them that appear across platforms, in order to understand the nature of the category of these features, and also to understand how they appear (Twitter, Text message, Facebook and Email); furthermore, this should provide some direction for future research. A full listing of all 275 features for each platform can be found in the Appendix B).

Further analyses have been conducted, and in this experiment, the top 10 stylometric features from each platform have been compared with other platforms in order to find the common features between platforms. In Table 5-1, light yellow represents the common features shared between the four platforms, which are the number of punctuation marks (syntactic feature) and frequency of missing an uppercase letter when starting a sentence (short messages feature), F52 and F231 respectively. While the light orange colour represents the common features shared between the three platforms, that is, number of punctuation marks (syntactic feature), number of uppercase characters (lexical), and number of

characters (lexical); F55, F3 and F1 respectively. The light green colour represents the common features shared between the two platforms, that is, the number of punctuation marks (syntactic feature), number of alphabets (lexical feature), number of special characters (lexical) and average sentence length in terms of characters (structure): F54, F2, F39, and F213 respectively.

Table 5-1: The top 10 stylometric features for the platforms (Twitter, SMS, Facebook and Email)

| Twitter | SMS | FB | Email |
|---|---|---|---|
| 31 | 27 | 52 | 29 |
| 231 | 232 | 55 | 38 |
| 55 | 231 | 54 | 50 |
| 3 | 52 | 1 | 55 |
| 1 | 209 | 2 | 39 |
| 2 | 215 | 231 | 102 |
| 52 | 233 | 213 | 51 |
| 54 | 274 | 212 | 52 |
| 213 | 1 | 3 | 231 |
| 39 | 3 | 214 | 42 |

4 Platforms
3 Platforms
2 Platforms

Table 5-1 demonstrates that common stylometric features are positively shared between platforms. It shows that the Twitter, Text message, Facebook and Email platforms share these features: F52 number of punctuation marks (syntactic feature) and F231 frequency of missing an uppercase letter when starting a sentence (short messages feature). While F55, number of punctuation marks (syntactic), is shared between Twitter, Facebook and Email. F3, number of uppercase characters (lexical), is shared between Twitter, Text message and Facebook. F1, number of characters (lexical), is shared between Twitter, Text message and Facebook. F54, number of punctuation marks (syntactic), is shared between Twitter and Facebook. F2, number of alphabets (lexical feature), is

shared between Twitter and Facebook. F39, number of special characters (lexical), is shared between Twitter and Email. Lastly, F213, average sentence length in terms of character (lexical), is shared between Twitter and Facebook.

As has been explained in the previous chapter, the best performance was achieved by the Text message and Emails platforms with an EER of 7.97% and 13.11% respectively, for the Train/Test ratio 70/30, with GB as the best classifier, with performance significantly increasing for the other two platforms with an EER of 20.16 % and 25% for Twitter and Facebook respectively. An analysis of the dataset shows, in terms of size and composition, that the SMS Text message and Email repositories represent either end of the spectrum, suggesting volume is less likely to be a determining factor over the composition of the message itself. The results show a significant difference in performance depending on the platform utilised. More importantly, it can be noted that Twitter and Facebook are similar and share features to some extent and do not match with the others; they are, F54, F2, and F213.

Table 6-2, 6-3 and 6-4 show the most common features according to their classification within the the top 10, top 20 and top 30 stylometric features. Table 6-2 focuses on the top 10 for population, whereas 6-3 and 6-4 focus on the top 20 and 30 for population.

Table 5-2 below shows that lexical features are covered on most platforms, since they appear five times, while it can also be noticed that syntactic features came second and appear three times, and structure and short message features each appear only once.

In general, it appears that populationally, lexical feature seems to play a larger role than other features across platforms. Lexical features appeared five times,

syntactic features appeared three times, and structure features appeared only once. Lexical features are the most common feature for population on multi-platforms, because they are involved in more than one platform. Table 5-2 below shows the output of the most common features when the first top ten features were captured and investigated.

Table 5-2: Results of the common features when the first top 10 features were captured for population across platforms

| Common features | | | Platforms | | | | No.Platforms |
|---|---|---|---|---|---|---|---|
| #features | Features | | Twitter | SMS | FB | Email | |
| F52 | # punctuation | (Syntactic) | √ (P7) | √(P4) | √(P1) | √(P8) | 4 |
| F231 | Frequency of missing an uppercase letter when starting a sentence | (short messages feature) | √(P2) | √(P3) | √(P6) | √(P9) | 4 |
| F55 | # punctuation | (Syntactic) | ¶ | - | ¶ | ¶ | 3 |
| F3 | # uppercase characters | (Lexical) | ¶ | ¶ | ¶ | - | 3 |
| F1 | # characters | (Lexical) | ¶ | ¶ | ¶ | - | 3 |
| F54 | # punctuation | (Syntactic) | § | - | § | - | 2 |
| F2 | # alphabets | (Lexical) | § | - | § | - | 2 |
| F39 | # special character | (Lexical) | § | - | - | § | 2 |
| 213 | Average sentence length in terms of character | (Structure) | § | - | § | - | 2 |
| 219 | #words with 5 characters | (Lexical) | § | § | - | - | 2 |

P= Position

√= Four platforms

¶= Three platforms

§= Two platforms

Having set out the results in Table 5-2, it is necessary to conduct an analysis of the dataset, which shows, in terms of size and composition, that for the Text message (position 4) and Facebook (position 1) platforms, the feature # punctuation seems strongest and may represent either end of the spectrum, suggesting volume is less likely to be a determining factor over the composition

of the message itself. For example, they shared the highest position in the ranking for # punctuation (syntactic, F52). On the other hand for another feature, an analysis of the dataset also shows, in terms of size and composition, that Twitter (position 2) and Text message (position 3) platforms represent either end of the spectrum, suggesting volume is less likely to be a determining factor over the composition of the message itself. For example, they are shared in feature frequency of missing an uppercase letter when starting a sentence (short messages feature, F231). This is the first study to attempt to solve the cross-platform author verification problem by exploiting some features to find the most discriminating features across modern messaging platforms for a population-base.

Further investigations have been conducted in order to investigate whether the syntactic, lexical, and short message features also exist in the top 20 among the population across platforms in order to discover whether lexical and syntactic features also exist in the top 20 among the population across platforms; therefore the top twenty have been analysed.

In this experiment, the top 20 stylometric features from each platform have been examined and compared with other platforms in order, firstly, to expand on the common features between platforms and, secondly, to investigate whether lexical features also exist in the top 20. In addition to the previous lexical features being common in the top 10, and the output of lexical features being more verifiable in terms of some common features, the number of top features was increased to twenty, and the selection of the top features that are shared between two or more platforms was explored. Table 5-3 below shows the results for other additional common stylometric features shared between platforms (a full listing of the top stylometric features can be found in Appendix B).

Table 5-3: Results of the common features when the first top 20 features were captured for population across platforms

| Common features | | | Platforms | | | | No.Platforms |
|---|---|---|---|---|---|---|---|
| #features | Features | | Twitter | SMS | FB | Email | |
| F3 | Number of uppercase characters | (Lexical) | √ (P4) | √(P10) | √(P9) | √(16) | 4 |
| F54 | Number of punctuation | (Syntactic) | √(P8) | √(15) | √(P3) | √(17) | 4 |
| F55 | Number of punctuation | (Syntactic) | √ | √ | √ | √ | 4 |
| F213 | Average sentence length in terms of character | (Structure) | √ | √ | √ | √ | 4 |
| F32 | Number of special character | (Lexical) | § | - | § | - | 2 |
| F48 | Number of special character | (Lexical) | ¶ | - | ¶ | ¶ | 3 |
| F210 | Total number of words | (Lexical) | ¶ | ¶ | ¶ | - | 3 |
| F214 | Average sentence length in terms of word | (Lexical) | ¶ | ¶ | ¶ | - | 3 |
| F27 | Number of alphabet a-z | (Lexical) | § | § | - | - | 2 |
| F212 | Average word length | (Lexical) | ¶ | | ¶ | ¶ | 3 |
| F227 | Number of words with more than 12 chars | (Structure) | § | - | - | § | 2 |
| F209 | Total number of sentences (Structure). | (Structure) | ¶ | ¶ | ¶ | - | 3 |
| F23 | Number of alphabet a-z | (Lexical) | § | - | § | - | 2 |
| F228 | Frequency of a smile face | (Emotional) | | ¶ | ¶ | ¶ | 3 |
| F51 | Number of punctuation | (Syntactic) | - | √ | - | √ | 2 |
| F2 | Number of alphabets | (Lexical) | ¶ | ¶ | ¶ | | 3 |
| F232 | Frequency of missing a period or other punctuation to end a sentence | (Short Messages feature) | - | § | § | - | 2 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| F8 | Number of alphabet a-z | (Lexical) | § | - | § | - | 2 |
| F22 | Number of alphabet a-z | (Lexical) | § | - | § | - | 2 |
| F233 | Frequency of missing the word "I" or "We" in a sentence | (Short Messages feature) | ¶ | ¶ | ¶ | - | 3 |

P= Position

In Table 5-3 above, it can be seen that when some features increased, they became common to other platforms, such as: F3, F54, F55, and F213. It is clear that lexical features took the lead and were ahead of the other features between platforms, even when the top 20 features were considered, as they covered most platforms and appeared eleven times. While it can also be noticed that syntactic features came second as they appeared three times.

In general, lexical and syntactic are the most common features across the Twitter, Text message, Facebook and Email platforms, even if the number of features increases to include the top 20 features. An analysis of the dataset shows that in terms of size and composition, the Twitter (position 4) and Facebook (position 9) platforms represent either end of the spectrum, suggesting volume is less likely to be a determining factor over the composition of the message itself; for example they are shared in # uppercase characters (lexical, F3). On the other hand, for another feature, an analysis of the dataset also shows, in terms of size and composition, that the Twitter (position 8) and Facebook (position 3) platforms represent either end of the spectrum, suggesting volume is less likely to be a determining factor over the composition of the message itself. For example, they are shared in # punctuation (Syntactic, F54).

Further investigations have been conducted, and in the next experiment, the top 30 stylometric features for each platform have been examined and compared with

other platforms in order, firstly, to expand on the common features between platforms and, secondly, to investigate whether the lexical features also exist in the top 30. In addition to the previous lexical features being common in the first top 10, 20 and the output of lexical features being more verifiable in terms of some common features, the number of top features increased to thirty. The reason for the top 30 features being taken to explore common features between platforms, is because they are the least top features that have achieved a good performance between four platforms. For example, the top 30 features achieved a good performance in Twitter platform. Consequently, it was treated as the least useful feature. In addition to the previous features, the number of features have been increased to the top 30 features. Table 5-4 below shows the results for other additional common stylometric features shared between platforms. The full listing of the top stylometric features can be found in Appendix G.

Table 5-4 Results of the common features when the first top 30 features were captured for population across platforms

| Common features | | | Platforms | | | | No.Platforms |
|---|---|---|---|---|---|---|---|
| #features | Features | | Twitter | SMS | FB | Email | |
| F212 | Average word length | (Lexical) | √(P16) | √(P22) | √(P8) | √(P14) | 4 |
| F1 | Number of characters | (Lexical) | √(P5) | √(P9) | √(P4) | √(P28) | 4 |
| F232 | Frequency of missing a period or other punctuation to end a sentence | (Short Message features) | ¶ | ¶ | ¶ | - | 3 |
| F8 | number of alphabet a-z | (Lexical) | ¶ | ¶ | ¶ | - | 3 |
| F233 | Frequency of missing the word “I” or “We” when starting a sentence | (Short Messages feature) | ¶ | ¶ | ¶ | - | 3 |
| F51 | Number of punctuation | (Syntactic) | ¶ | ¶ | | ¶ | 3 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| F211 | Number of short words (less than four characters) | (Lexical) | ¶ | ¶ | ¶ | - | 3 |
| F4 | Number of alphabet a-z | (Lexical) | § | | § | - | 2 |
| F236 | Number of Face With Tears of Joy 😂 | (Emotional) | - | § | § | - | 2 |
| F23 | Number of alphabet a-z | (Lexical) | - | § | § | - | 2 |
| F58 | Number of punctuation | (Syntactic) | - | § | § | - | 2 |
| F56 | Number of punctuation | (Syntactic) | - | § | - | § | 2 |
| F27 | Number of alphabet a-z | (Lexical) | ¶ | ¶ | - | ¶ | 3 |

P= Position

In Table 5-4 , it can be noticed that when some features were increased to more than ten features, they became common to other platforms such as: F212, F1, 232, F8, F51, F23 and F233. It is clear that lexical features took the lead and were ahead of the other features between platforms, even when the top 30 features were considered, as they covered most platforms and appeared seven times. While it can also be noticed that syntactic features came second as they appeared three times. While, short message features appeared twice, and features of emotional icon appeared only once.

All in all, the most common features between these platforms were lexical, even if the number of features increased from the top 10 features, to the top 20 features, through to the top 30 features.

An analysis of the dataset shows that in terms of size and composition, when the first top 30 stylometric features for platform were examined, the Facebook (position 8) and Email (position 14) platforms represent either end of the spectrum, suggesting volume is less likely to be a determining factor over the composition of the message itself. For example, they are shared in average word

length (lexical feature). On the other hand, for another feature, an analysis of the dataset shows that in terms of size and composition when the first top 30 stylometric features for platform were examined, Facebook (position 4) and Twitter (position 5) represent either end of the spectrum, suggesting volume is less likely to be a determining factor over the composition of the message itself. For example, they are shared in #characters (Lexical). Having said that, the reason for selecting the top-most features is because it should show that there are some common features across the first 10, 20 and 30 features. Table 5-5 below summarises the most common features that worked with all four platforms, including the categories, based on exploring the top 30 features.

Table 5-5: Results of the top common features including categories between the four platforms

| Common features | | | Platforms | | | |
|---|---|---|---|---|---|---|
| #features | Features | | Twitter | SMS | FB | Email |
| F52 | Number of punctuation | (Syntactic) | √ | √ | √ | √ |
| F231 | Frequency of missing an uppercase letter when starting a sentence | (Short messages feature) | √ | √ | √ | √ |
| F3 | Number of uppercase characters | (Lexical) | √ | √ | √ | √ |
| F54 | Number of punctuation | (Syntactic) | √ | √ | √ | √ |
| F55 | Number of punctuation | (Syntactic) | √ | √ | √ | √ |
| F213 | Average sentence length in terms of character | (Structure) | √ | √ | √ | √ |
| F212 | Average word length | (Lexical) | √ | √ | √ | √ |
| F1 | Number of characters | (Lexical) | √ | √ | √ | √ |

In general, lexical and syntactic are the most common features across the modern platforms of Twitter, Text message, Facebook and Email, even if the number of features increases to include the top 20 and top 30 features. This is the first study to explore the cross-platform author verification problem by exploiting the lexical

and syntactic features as the most discriminating features across modern messaging platforms using population-based.

### 5.2.2 User-Based Analysis (Common Feature Vectors that are User-Based)

The results of this experiment (as illustrated in Table 4-13) are encouraging and show that there are some strong features that could be further investigated concerning the feature analysis of individual based features across platforms. The experimental results of the classification algorithms have revealed that the stylometry of authors on platforms can be identified with a high degree of recognition (Abbasi, 2008). Therefore, feature analysis and the common features of authors have been explored. Since some authors have two, three or four platforms, identifying the common features of authors on different platforms was conducted based on their platform availability. Authors' dataset availability were divided into four platforms, three platforms and two platforms. The following sections provide examples and an analysis of common features for Authors who have four, three and two platforms.

- Feature Analysis for Authors with Four Platforms

As previously highlighted, some authors have four platforms. Table 5-6 shows the feature analysis for authors that have four platforms (i.e. Authors 1, 15, and 18 across platforms); the reason for selecting these users with four platforms is because the results of their EERs are not high or low across the four platforms (this has been assumed to be somewhat the average case between most platforms). The EER for each individual user, based on the user-based experimental results in Table 4-13, is provided (a full listing of all users' EERs for each individual platform can be found in Appendix K). For example, *Author 1*'s performance is: 5.5%, 17.8%, 20.8% and 10%, for SMS Text message, Twitter,

Facebook and Email respectively. The following table shows the results of the feature analysis for authors.

Table 5-6: Common features for users of platforms

**Author 1**

Platforms

| 1 | 2 | 3 | 4 |
|---|---|---|---|
| Tw | SMS | FB | Email |
| 55 | 28 | 1 | 1 |
| 214 | 1 | 55 | 224 |
| 53 | 233 | 13 | 213 |
| 215 | 232 | 9 | 52 |
| 1 | 275 | 210 | 49 |
| 40 | 2 | 214 | 53 |
| 2 | 210 | 215 | 229 |
| 3 | 234 | 2 | 44 |
| 58 | 229 | 213 | 43 |
| 227 | 55 | 16 | 40 |
| 4 | 53 | 3 | 18 |
| 23 | 3 | 22 | 12 |

**Author15**

Platforms

| 1 | 2 | 3 | 4 |
|---|---|---|---|
| Tw | SMS | FB | Email |
| 1 | 1 | 1 | 1 |
| 32 | 28 | 213 | 40 |
| 2 | 3 | 224 | 20 |
| 33 | 214 | 275 | 213 |
| 55 | 213 | 9 | 227 |
| 3 | 220 | 22 | 30 |
| 24 | 2 | 214 | 228 |
| 4 | 4 | 3 | 103 |
| 19 | 24 | 2 | 59 |
| 224 | 211 | 219 | 214 |
| 20 | 12 | 29 | 49 |
| 56 | 8 | 33 | 4 |

**Author 18**

Platforms

| 1 | 2 | 3 | 4 |
|---|---|---|---|
| Tw | SMS | FB | Email |
| 1 | 1 | 1 | 1 |
| 32 | 52 | 233 | 44 |
| 2 | 28 | 55 | 49 |
| 3 | 233 | 4 | 43 |
| 213 | 211 | 229 | 210 |
| 211 | 2 | 213 | 56 |
| 22 | 215 | 215 | 29 |
| 214 | 3 | 218 | 22 |
| 23 | 212 | 214 | 234 |
| 4 | 237 | 216 | 53 |
| 7 | 12 | 9 | 8 |
| 5 | 55 | 212 | 223 |

As shown in Table 5-6, there is a clear indication that there are a set of common stylometric features shared between platforms by the authors. The top thirty features were captured for authors after ranking them using the RF algorithm, and dealing with this as a two-class classification problem (see Appendix H). For example, *Author 1* shares common features on four platforms (F1, F53, F234, F4). In addition, there are common features between three platforms (F55, F214, F2, F3, F210.. etc.). There are also common features between two platforms (F215, F40, F29, F233..etc.). The same procedure applies to all users, and Table 5-7 illustrates some sets of common features (patterns) shared between platforms by the Authors, and these are coded in colour to differentiate the order of platforms between each other (a full listing of user features for each platform is provided in Appendix H).

Table 5-7: Common features for users who have 4 platforms

**Common features (e.g. User 1)**

| Features | Platforms | Type |
|---|---|---|
| F1 | 1,2,3,4 | **Lexical** |
| F53 | 1,2,3,4 | Syntactic |
| F55 | 1,2,3 | Syntactic |
| F214 | 1,3,4 | **Lexical** |
| F215 | 1,3 | **Lexical** |
| F40 | 1,4 | **Lexical** |
| F2 | 1,2,3 | **Lexical** |
| F3 | 1,2,3 | **Lexical** |
| F23 | 1,3 | **Lexical** |
| F210 | 1,2,3 | **Lexical** |
| F213 | 1,3,4 | **Lexical** |
| F29 | 1,3 | **Lexical** |
| F22 | 1,3 | **Lexical** |
| F234 | 1,2,3,4 | Emotion |
| F211 | 1,3 | **Lexical** |
| F19 | 1,3 | **Lexical** |
| F18 | 1,4 | **Lexical** |
| F229 | 2,4 | **Lexical** |
| F218 | 3,4 | **Lexical** |
| F220 | 1,4 | **Lexical** |
| F24 | 1,3 | **Lexical** |
| F16 | 1,3,4 | **Lexical** |

**Common features (e.g.User 15)**

| Features | Platforms | Type |
|---|---|---|
| F1 | 1,2,3,4 | **Lexical** |
| F2 | 1,2,3,4 | **Lexical** |
| F33 | 1,3 | **Lexical** |
| F3 | 1,2,3 | **Lexical** |
| F24 | 1,2,3 | **Lexical** |
| F4 | 1,2,3,4 | **Lexical** |
| F19 | 1,3 | **Lexical** |
| F224 | 1,3 | **Lexical** |
| F20 | 1,4 | **Lexical** |
| F214 | 1,2,3,4 | **Lexical** |
| F213 | 1,2,3,4 | **Lexical** |
| F7 | 1,3 | **Lexical** |
| F215 | 1,3 | **Lexical** |
| F40 | 1,4 | **Lexical** |
| F5 | 1,3 | **Lexical** |
| F211 | 1,2,3 | **Lexical** |
| F218 | 1,3 | **Lexical** |
| F9 | 1,3 | **Lexical** |
| F212 | 1,2,3 | **Lexical** |
| F13 | 1,3 | **Lexical** |
| F16 | 1,3 | **Lexical** |
| F28 | 2,3 | **Lexical** |

**Common features (e.g User 18)**

| Features | Platforms | Type |
|---|---|---|
| F1 | 1,2,3,4 | **Lexical** |
| F2 | 1,2,3 | **Lexical** |
| F3 | 1,2,3 | **Lexical** |
| F213 | 1,3 | **Lexical** |
| F211 | 1,2,3 | **Lexical** |
| F214 | 1,3,4 | **Lexical** |
| F4 | 1,3 | **Lexical** |
| F13 | 1,3 | **Lexical** |
| F9 | 1,3 | **Lexical** |
| F25 | 1,3 | **Lexical** |
| F218 | 1,3 | **Lexical** |
| F233 | 2,3 | **Social** |
| F215 | 2,3,4 | **Lexical** |
| F212 | 2,3 | **Lexical** |
| F55 | 2,3 | **Syntactic** |
| F22 | 1,3,4 | **Lexical** |
| F23 | 1,3 | **Lexical** |
| F8 | 3,4 | **Lexical** |
| F10 | 3,4 | **Lexical** |
| F234 | 3,4 | **Emotion** |
| F29 | 3,4 | **Lexical** |
| F56 | 1,4 | **Syntactic** |

| |
|---|
| Platform 1=Twitter |
| Platform 2=SMS |
| Platform 3=Facebook |
| Platform 4=Email |

Table 5-7 shows that there are some strong common features that can be used for feature analysis to build a sufficient user profile across platforms, for example *Author 15*. Although there are similarities concerning some common types of features among authors, for example, lexical features, the features used often vary among users. Interestingly, most users have specific categories within these types that are different from others, and this makes them common to them. For example, in Table 5-8 below, *Author 1* is distinguished by their common features, such as F53 and F229, which are distinctive and differ from the other authors, so it may be possible to build a linguistic profile of that user. In addition, it can be noted that feature F53 is robust and distinct for *Author1* and shared on all four

platforms, but not robust for others, and it does not exist even within their features on their platforms. Therefore, these differences can make subsets and common features for *Author1* and thus distinguish and identify the user from others based on their different common features. The same applies to other users.

This is an interesting outcome across the analysis of the platforms, in that when a multi-platform is used, there are robust features of users shared and commonalities between these platforms, as indicated by, for example, F229 shared between SMS text message and Email, and F218 shared between Facebook and Email for *Author* 1. Further analysis shows that when the author used only two platforms, if these platforms are integrated with other platforms, these features may not appear. This can distinguish the author if he or she uses two platforms, as when investigating a suspect, the suspect often tries to hide his platform, but if there is insufficient information or a lack of platforms, it is possible to depend on the features on the available platforms, as this may lead to finding specific common features of the suspect regardless of what is shared between other platforms. However, the higher the number of platforms, the easier it is to verify and identify the user, and vice versa, as it is possible to obtain different unique features shared between the available platforms.

An analysis of *User1, User15* and *User 18*'s dataset shows that in terms of size and composition, the Text message and Email platforms represent either end of the spectrum, suggesting volume is less likely to be a determining factor over the composition of the message itself. For example, lexical feature seems to play a larger role than other features across platforms. Lexical features are the most common features for users who have four platforms, because they are involved in more than one platform. Table 5-8 below shows the differences in common features between users.

Table 5-8: Common features for users who have four platforms

| Common features | | Platforms | | | |
|---|---|---|---|---|---|
| user | Features | Twitter | SMS | Facebook | Email |
| **User 1** | F1,**F53**,F234,F4 | **√** | **√** | **√** | **√** |
| | F55,F2,F3,F210, | ¶ | ¶ | ¶ | - |
| | F214,F213 | ¶ | - | ¶ | ¶ |
| | F215,F23,F29,F22,F211,F19,F5,F9,F24 | § | - | § | - |
| | F40,F18,F220 | § | - | - | § |
| | F233 | § | - | § | - |
| | **F229** | - | § | - | § |
| | **F218** | - | - | § | § |
| **User 15** | F1,F2,F4,F214,F213 | √ | √ | √ | √ |
| | F33,F19,F224,F7,F215,F5,F218,F9,F13,F16 | √ | - | √ | - |
| | F3,F23,F211,F212 | ¶ | ¶ | ¶ | - |
| | **F20**,F40 | § | - | - | § |
| | F28 | - | § | § | - |
| | F12 | - | ¶ | ¶ | ¶ |
| | F59 | - | § | - | § |
| **User 18** | F1 | √ | √ | √ | √ |
| | F2,F3,F211 | ¶ | ¶ | ¶ | - |
| | F213,F4,F13,F9,F25,F218,**F23** | § | - | § | - |
| | F233,F212,F55, | - | § | § | - |
| | 214,F22 | ¶ | - | ¶ | ¶ |
| | F215 | - | ¶ | ¶ | ¶ |
| | F8,F10,F234, F29 | - | - | § | § |
| | F56 | § | - | - | § |

- Feature Analysis for Users who have Three Platforms

There are number of common stylometric features shared between authors who have three platforms. For example, *Author 21* shares common features across three platforms (e.g. F1, F214, F215, F3. etc). In addition, there are also common features across two platforms (e.g. F2, F211, F9, F53, F24. etc.), and the same applies to all authors. The reason for selecting these users is because the result of their EERs is somewhat not high and not low across the platforms (it has been assumed to be the average case between most platforms). The EER for each individually user, based on the user-based experimental results of Table 4-13, is provided (a full listing of all users' EERs for each individual platform can be found in Appendix K). Table 5-9 below illustrates some of the common features (patterns) shared between platforms by authors, and these are colour coded to

differentiate the order of platforms between each other (a full listing of user features for each platform is provided in Appendix H).

Table 5-9: Common features for users who have three platforms

**e.g (User 21)**

| Feature | Platforms | Type |
|---|---|---|
| F1 | 1,3,4 | Lexical |
| F214 | 1,3,4 | Lexical |
| F215 | 1,3,4 | Lexical |
| F3 | 1,3,4 | Lexical |
| F217 | 1,3,4 | Lexical |
| F2 | 1,3 | Lexical |
| F212 | 1,3,4 | Lexical |
| F211 | 1,3 | Lexical |
| F9 | 1,3 | Lexical |
| F53 | 1,3 | Syntactic |
| F232 | 1,3 | Social |
| F4 | 1,3,4 | Lexical |

**e.g (User 25)**

| Feature | Platforms | Type |
|---|---|---|
| F1 | 2,3,4 | Lexical |
| F214 | 2,3 | Lexical |
| F213 | 2,3,4 | Lexical |
| F22 | 2,3 | Lexical |
| F217 | 2,3 | Lexical |
| F215 | 2,3 | Lexical |
| F3 | 2,3,4 | Lexical |
| F212 | 3,4 | Lexical |
| F9 | 3,4 | Lexical |
| F5 | 3,4 | Lexical |
| F29 | 3,4 | Lexical |
| F2 | 2,3,4 | Lexical |

**e.g (User 30)**

| Features | Platforms | Type |
|---|---|---|
| F1 | 1,2,4 | Lexical |
| F53 | 1,2 | Syntactic |
| F17 | 1,4 | Lexical |
| F213 | 1,2,4 | Lexical |
| F3 | 1,2 | Lexical |
| F2 | 1,2,4 | Lexical |
| F4 | 1,2,4 | Lexical |
| F220 | 1,4 | Lexical |
| F59 | 1,4 | Syntactic |
| F22 | 1,2 | Lexical |
| F214 | 1,2,4 | Lexical |
| F215 | 1,2,4 | Lexical |

| |
|---|
| Platform 2=SMS |
| Platform 3=Facebook |
| Platform 4=Email |

Table 5-9 shows there are some strong common features, for example lexical features can be used for feature analysis to build a sufficient user profile across platforms, such as for *User 25*. Despite the similarity of some common types of features among authors, such as lexical features, the common categories used and which are part of this type, vary among authors. Most authors have specific features within these types that are different from others, and this makes them common to them. Although there are several strong features that can be differentiated between authors, this may be one of the limitations if the number of authors increases. For example, in the following table it can be seen that *User 21* can be distinguished by some of their common categories, which are different from others, such as F24 and F33. Therefore, these differences are common features for *User 21* and distinguish *User 21* from others. The same applies to other users. Table 5-10 below shows the differences in features between users,

including the common features for each author, and Figure 5-1 shows an example of a set of features for *User 21* that are common to all available platforms.

Table 5-10: Common features for users who have 3 platforms

| Common features | | Platforms | | | |
|---|---|---|---|---|---|
| user | Features | Twitter | SMS | Facebook | Email |
| User 21 | F1,F214,F215,F3,F217,F212,F4,F19,F213 | √ | - | √ | √ |
| | F2,F211,F9,F53,F232 | √ | - | √ | - |
| | **F24** | √ | - | - | √ |
| | **F33** | √ | - | √ | - |
| | F13 | - | - | √ | √ |
| User 25 | F1,F213,F3,F2 | - | √ | √ | √ |
| | F214,F22,F217,F215 | - | √ | √ | - |
| | F212,**F9**,**F5**,F29 | - | - | √ | √ |
| User 30 | F1,F213,F2,F4,F214,F215 | √ | √ | - | √ |
| | F53,F3,F22 | √ | √ | - | - |
| | **F17**,**F220**,F59,F12 | √ | - | - | √ |
| | F210,F234 | - | √ | - | √ |

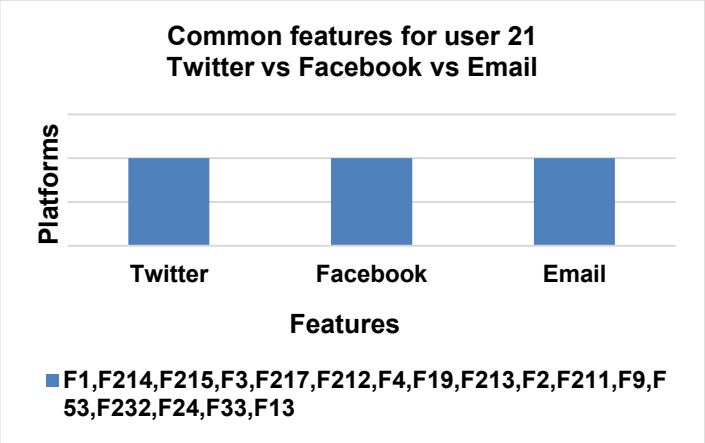



Figure 5-1 : A set of common features for user in multiplatforms

In general, an analysis of *User 25* and *User 30's* dataset shows that in terms of size and composition, Text message and Email repositories represent either end of the spectrum, suggesting volume is less likely to be a determining factor over the composition of the message itself. For example, lexical features seem to play a larger role than other features across platforms. Lexical features are the most

common feature for users who have three platforms, because they are involved in more than one platform.

- Feature Analysis for Users who have Two Platforms

There are also a number of common stylometric features shared between two platforms for authors. For example, *Author 48* shares common features on two platforms (F1, F2, F213, F3..etc).The same procedure applies to all users' features. In general, lexical features seem to play a larger role than other features across platforms. Lexical features are the most common for users who have two platforms, because they are involved in more than one platform. Table 5-11 below demonstrates some common features shared between two platforms for some authors, and shows a set of common lexical features for *Author 48* on all available platforms (a full listing of all the top user features for each platform for a user as an example is provided in the Appendix section).

Table 5-11: Common features for users who have two platforms

**e.g (User 48)**

| Feature | Platforms | Type |
|---|---|---|
| F1 | 1,3 | Lexical |
| F2 | 1,3 | Lexical |
| F213 | 1,3 | Lexical |
| F3 | 1,3 | Lexical |
| F217 | 1,3 | Lexical |
| F214 | 1,3 | Lexical |
| F212 | 1,3 | Lexical |
| F19 | 1,3 | Lexical |

**e.g (User 25)**

| Features | Platforms | Type |
|---|---|---|
| F1 | 3,4 | Lexical |
| F213 | 3,4 | Lexical |
| F2 | 3,4 | Lexical |
| F214 | 3,4 | Lexical |
| F3 | 3,4 | Lexical |
| F5 | 3,4 | Lexical |
| F4 | 3,4 | Lexical |
| F24 | 3,4 | Lexical |

| |
|---|
| Platform 1=Twitter |
| Platform 2=SMS |
| Platform 3=Facebook |
| Platform 4=Email |

## 5.3 Unified Feature Profile

The previous section provided some of the basis for understanding that there appears to be commonality between some features. This section will further

investigate the exploration of messages and features to understand whether a unified profile could be created to enable identity verification across platforms, and the extent to which stylometric features are common when a user's file is unified. Through this approach, it is possible to assess the ability of feature vectors to verify the identity of a user on dependant platforms.

This is the first study to attempt to address the cross-platform author verification problem by exploiting certain features across modern messaging platforms. This approach has basically involved bringing together as much data as possible from the profiles that are available across modern platform to give the best result. It has included identifying as much genuine information from users as possible and incorporating this into a classifier; this approach has then suggested what factors can possibly be picked up.

The aim of this section is to explore common features on the platforms for verifying the authors of different platform text samples. The method involves verifying the authorship of different text samples by unifying all features from different messaging platforms, for example unifying the most discriminating features found on Twitter, Text message, Facebook and Email. The main aim is:

- To find a systematic and forensically automated method to be used under a one umbrella mechanism to assist linguistic experts to create profiles of authors flexibly and reliably cross platforms.
- To assist forensic experts in identifying the movements of features that may contribute towards tracking an author’s profile across platforms (identifying proper features across platforms can lead to identifying the author).

- To support intelligence applications to analyse aggressive and threatening messages.
- To show, empirically, the movement of user profiles across platforms, and also to show an understanding of the impact of unifying the most discriminating features of users on multi-platforms.
- To find different ways of unifying subsets of common user features across platforms, which should lead ultimately to identifying the author (explore how a profile can be unified across platforms), as well as including an investigation into which features across these platforms prove to be discriminative and useful (how similar they are across platforms).
- Finally, to study a possible way of finding common features by unifying user features to verifying the author. In this novel method, by unifying the top discriminating features of an author, it may be possible to conduct user profile verification analysis cross platforms. Thus, the historical datasets collected from four different messaging systems have been examined.

The methodology used for the data collection and pre-processing of these platform datasets is described in detail in Chapter Four, and population and user based feature analysis approaches were examined in the previous section. The unified feature-based user verification approach across platforms is investigated in this section.

### 5.3.1 Methodology for the Unified User-Based Verification Approach

Exploring unified profiling techniques requires a unified author-based feature set for all platform datasets. This unified profiling approach is based on the individual's most discriminating features to form a profiling template, and so it was necessary to verify the unified users' most discriminating and most important feature profiles. As illustrated earlier, the reason for verifying the user's unified

top discriminating individual features for different text messages is based on unifying individual linguistic user profiles across platforms to create a user profile model.

A set of experiments were conducted with different settings for verifying a given user's different samples.

- In a unfied user-based approach, the dataset which contains all users' various types of text samples (i.e Twitter, Facebook, Text message, and Email samples) are put through a process to extract features. The process was used to extract the features of each user in order to prioritise them for identifying only the most relevant user features. Therefore, the RF deals with this as a two-class classification problem.
- This included selecting a varying number of features. The reason for ordering 10, 20, 30, 50, 100 and 275 stylometric features is because they have been shown to produce the best results for both population and user-based feature profiles.
- In terms of user data modelling, as mentioned in the previous chapter, no single classification method can solve all classification problems; however, three different classification algorithms were used to find the optimum classifier to verify message authorship, which included SVM, RF and GB. The GB was found to be the optimal classifier for the best performance in this research for the previous experiments - both population and user based. Each classifier was tested using a different set of features, as presented in the next section.

In order to reach the desired result, with the help of the historical datasets available for this research (Text messages, Twitter tweets, Facebook posts and

Email messages), for all four platform datasets, each classifier was tested in one setting of the Train/Test ratio 70/30. This is because this setting has been revealed to be the best setting for population and user-based techniques from among other settings, each of which was repeated by the number of authors in the dataset by using a one-vs-all approach. In order to achieve the desired goal and to understand the most important features for the user-base, the total number of experiments that have been applied are as follows: For all datasets, the total number of experiments which equals 480 tests, as shown in Table 5-12.

Table 5-12: Total number of tests for all datasets

| Platform | #Users | #Classifiers | Configuration | Total experiments |
|---|---|---|---|---|
| Twitter | 41 | 3 | 1 | 123 |
| Text message | 26 | 3 | 1 | 78 |
| Facebook | 46 | 3 | 1 | 138 |
| Email | 47 | 3 | 1 | 141 |
| | | | *Total | 480 |

The experiments were based on 275 features sets, and five different types of linguistic features sets, as lexical features (character and word-based features), syntactic features, structural features, short message features and emotional features respectively. The type of stylometric feature sets, and how they were selected, is described in detail in Chapter Four.

### 5.3.2 Experimental Results

The preliminary test combined the most influential feature analysis of the author and examined whether it could possibly be unified in order to explore common features, and to investigate lexical features. Moreover, the most discriminating features across platforms may be useful if they are automated so as to combine the most discriminative features automatically. As illustrated earlier, this experiment explored the impact of automated features to achieve the following objectives: First, to understand the underlying dataset to determine whether there

are unique patterns that can be used to discriminate individuals. Second, to assist in identifying the movement of features that may contribute towards and track the subset of features of a suspect/an author profile across platforms (identifying proper features leads to identifying the author). Third, to explore all possible ways of identifying the subset of common user features across platforms, which should lead ultimately to identifying the common features of an author (explore how a profile can be used across platforms). Table 5-13 below shows the results from verifying the authorship of a given number of unified text messages from different platforms.

Table 5-13: Unified platform model

| Test ID | Train/Test ratio | Feature tested | Performance EER (%) | | |
|---|---|---|---|---|---|
| | | | SVM | GB | RF |
| Test 1 | 70/30 | Top 10 | 14.91 | 10.78 | 11.27 |
| Test 2 | | Top 20 | 14.34 | 9.76 | 11.05 |
| Test 3 | | Top 30 | 14.49 | 9.61 | 11.02 |
| Test 4 | | Top 50 | 14.26 | 9.49 | 11.32 |
| Test 5 | | Top 100 | 15.81 | 9.46 | 12.06 |
| Test 6 | | All | 18.51 | 9.47 | 13.64 |

Table 5-13 shows the results of verifying the unified most discriminating features for platforms (Twitter, Text message, Facebook and Email) using all 275 stylometric features. The top 100 produced the best performance and yielded an EER of 9.46%. However, the unified feature model's performance is supposed to be worse than the user based experimental results, since for user-based verification, every platform is treated individually. As described earlier in Chapter Five, the previous performance for the verification of author-based features across platforms showed the following EER: Twitter, Text message, Facebook and Email: 20.28%, 7.97%, 23.78 and 12.03% respectively. While in the unifying experiment, the performance achieved an EER of 9.46%.

Across all four platforms, a performance 9.46% was achieved. In addition, this work has technically improved upon using isolated individual platforms, as the results seem to suggest that this approach is better. Furthermore, this approach has not been suggested or used before. Although these results need to be further researched, they are positive, and suggest that it is possible to reach even better performance in the future.

More importantly, a minimum amount of data is required to achieve reasonable performance. In addition, if a system that uses four platforms is introduced, obtaining data from those four platforms should ensure meeting the minimum requirements, although it is more difficult than using one platform. However, and more importantly, from a pragmatic perspective, using a single classifier approach is pragmatically better than using an individual approach.

For individual user performance, since this research requires knowing as much as possible about what the most common features of all platforms are, and to investigate the lexical features and whether this works when the platforms are analysed (based on the success in the previous experiment), it was decided to select users who have used all four unified platforms. For example, *User 1* achieved an EER of 6.5%, *User 15* achieved an EER of 9.1%, and *User 18* achieved an EER of 7.8%.

shows the top features distribution for user level unified performance across platforms, and *User 1*, *User 15* and *User 18* are coded in colour. The EER for the individual results show that there is a clear indication that there are a set of common stylometric features shared between platforms by authors, because the performance results achieved better performance. To illustrate the results of the

unified features for *User 1, User 15, User 18*, Table 6-14 below shows the unified features of these users.

Table 5-14: Authors' EER for unified platform model (Top 100)

| Users | EER | Users | EER |
|---|---|---|---|
| **1** | 6.5 | **26** | 9.1 |
| **2** | 9.2 | **27** | 2.4 |
| **3** | 8.4 | **28** | 11.8 |
| **4** | 14.2 | **29** | 4.1 |
| **5** | 13.5 | **30** | 6.8 |
| **6** | 15.0 | **31** | 6.4 |
| **7** | 14.6 | **32** | 5.8 |
| **8** | 14.2 | **33** | 10.7 |
| **9** | 17.1 | **34** | 5.7 |
| **10** | 14.1 | **35** | 6.5 |
| **11** | 4.6 | **36** | 11.0 |
| **12** | 7.8 | **37** | 10.3 |
| **13** | 11.6 | **38** | 7.0 |
| **14** | 16.1 | **39** | 15.8 |
| **15** | 9.1 | **40** | 6.1 |
| **16** | 9.9 | **41** | 10.5 |
| **17** | 14.3 | **42** | 10.1 |
| **18** | 7.8 | **43** | 6.5 |
| **19** | 11.1 | **44** | 5.6 |
| **20** | 5.9 | **45** | 7.1 |
| **21** | 6.4 | **46** | 8.1 |
| **22** | 11.1 | **47** | 9.5 |
| **23** | 7.4 | **48** | 5.7 |
| **24** | 12.6 | **49** | 10.6 |
| **25** | 13.1 | **50** | 5.4 |

Table 5-14 shows that some authors performed well, whereas some authors showed poor performance. More than half the authors in the dataset achieved an EER of less than 10%. Having said that, it was decided to select authors who have used all four unified platforms. The reason for selecting these users that have four platforms is because the result of their EERs is somewhat not high and not low across the four and unified platforms. For example, User 1 achieved an EER of 6.5%, User 15 achieved an EER of 9.1%, and User 18 achieved an EER of 7.8%. Compared to the EER of individual platforms, the performance of User 1 on individual platforms was as follows: Text message 5.6%, Email 10%, Twitter

17.8% and Facebook 20.9%, as shown in Chapter Five Table 4-14 User-Based. If it is assumed that calculating the total EERs for all the individual platforms' EER results and dividing them by the number of platforms, which are four platforms, will achieve the total result of EER, this leads to 13.57% for *User 1*, while the EER for *User 1* on this unified platform, as shown above in Table 6-14, is 6.5%. The same issue occurred for *User 15,* as for the individual platforms they achieved an EER of 0% for Text message, 18.2% for Twitter, 25.3% for Email and 30.2% for Facebook. Whereas calculating the total EER for all the individual platforms and dividing them by the number of platforms, which are four platforms, achieved a total result of an EER of 13.57%; while the EER result using this unified platform approach for *User 15* is 9.1%. The same is true for *User 18,* as on individual platforms they achieved an EER of 0% for Twitter, 7.3% for Text message, 9.7% for Email and 16.5% for Facebook. Whereas calculating the total EERs for all the individual platforms and dividing them by the number of platforms, which are four platforms, achieved an EER of 13.57%; while the EER using this approach for *User 18* was 7.8%, as shown in Table 6-14.

The above points indicate the following assumptions:

- The results of the performance of the unified platform model are better than the performance of individual platforms, even if the total EERs on individual platforms were calculated and divided by the number of platforms. For example, the performance of *User 1* on individual platforms was as follows: Text message 5.6%, Email 10%, Twitter 17.8% and Facebook 20.9%, as shown in Chapter Five Table 4-14 User-Based. If it is assumed that calculating the total EERs for all the individual platforms' EER results and dividing them by the number of platforms, which are four platforms, will achieve the total result of the EER, which leads to 13.57%,

while the EER on this unified platform, as above in Table 6-14 , is 6.5%. This gives a clear indication that the unified method can give more positive results than individual performance.

- The rank of features on the unified platforms is more effective than the rank of features on the single platform because of the presence of another platform.
- The existence of common linguistic characteristics on the united platform model is greater than the linguistic characteristics on independent platforms because the variety of features play a major role if they cross over a platform, and this helps to facilitate the verification process because there are often more similar characteristics if they are united.

Unified user features profiling analysis was used to select an effective subset of unified features for individual authors across platforms, as well as to explore the impact of unified discriminating features on multi-platforms. Table 5-15 below shows the results of the automated unified individual features for four different platforms.

Table 5-15: Users' unified features for platforms

| Unified user features | | |
|---|---|---|
| User 1 | User 15 | User 18 |
| feature | feature | feature |
| F28 | F1 | F1 |
| F1 | F32 | F52 |
| F233 | F4 | F28 |
| F232 | F3 | F233 |
| F275 | F2 | F32 |
| F2 | F213 | F213 |
| F53 | F20 | F237 |
| F210 | F224 | F55 |
| F55 | F214 | F212 |
| F229 | F28 | F4 |

| User 1 Platforms | | | |
|---|---|---|---|
| Tw | SMS | FB | E |
| F55 | F28 | F1 | F1 |
| F214 | F1 | F55 | F224 |
| 53 | F233 | F13 | F213 |
| F215 | F232 | F9 | F52 |
| F1 | F275 | F210 | F49 |
| F40 | F2 | F214 | F53 |
| F2 | F210 | F215 | F229 |
| F3 | F234 | F2 | F44 |
| F58 | F229 | F213 | F43 |
| F227 | F55 | F16 | F40 |
| F4 | F53 | F3 | F18 |

| User 15 Platforms | | | |
|---|---|---|---|
| Tw | SMS | FB | E |
| F1 | F1 | F1 | F1 |
| F32 | F28 | F213 | F40 |
| F2 | F3 | F224 | F20 |
| F33 | 214 | F275 | F213 |
| F55 | 213 | F9 | F227 |
| F3 | F220 | F22 | F30 |
| F24 | F2 | F214 | F228 |
| F4 | F4 | F3 | F103 |
| F19 | F24 | F2 | F59 |
| F224 | F211 | F219 | F214 |
| F20 | F12 | F29 | F49 |

| User 18 Platforms | | | |
|---|---|---|---|
| Tw | SMS | FB | E |
| F1 | F1 | F1 | F1 |
| F32 | F52 | 233 | F44 |
| F2 | F28 | F55 | F49 |
| F3 | F233 | F4 | F43 |
| F213 | F211 | F229 | F210 |
| F211 | F2 | F213 | F56 |
| F22 | F215 | F215 | F29 |
| 214 | F3 | F218 | F22 |
| F23 | F212 | F214 | F234 |
| F4 | F237 | F216 | F53 |
| F7 | F12 | F9 | F8 |

Table 5-15 shows the results of unifying the most discriminating features for *User1*, *User 15* and *User 18*. As can be seen, the left-hand columns presents the unified users' features and the right-hand columns present the users' features on their individual platforms. To understand the nature of the unification of features on the four platforms, three authors have been chosen, as they each used four platforms.

For *User* 1, it can be observed that feature 28 (lexical) was the first feature to appear as a unified feature on the table, which suggests that it may be robust for the user; in addition, it can be noticed that this feature was also the first feature for *User* 1 on the Text message platform. The same mechanism for features F32 (lexical) for *User 15* was the second robust features after feature 1 in the unified features. As it plays a major role for this user on the Twitter platform, this gives an indication that this user may use it continuously on the Twitter platform individually. The same technique applies to feature 52 for *User 18*, with the second robust feature when unified while it plays a major role for this user on the Text message platform, suggesting that this user may potentially always use it.

On the other hand, it can be observed that for *Users 15* and *18,* feature 1 (lexical) is a common feature, and it has appeared at the top of the unified features across the platform model, as well as appearing in the top features on all their individual platforms. Thus, these features can be investigated and may be common and robust features of those authors across platforms. It can also be proposed that F28 provides unique patterns that can be investigated to discriminate *User 1* from individuals on the Text message platform, while F1 shows a common feature for *Users 15* and *18*.

Indeed, selecting an effective or an optimum set of features is a critical and significantly important process because it will subsequently affect pattern

classification and the system's performance (Nguyen & Torre, 2010). These experiences and observations could contribute towards automating (user feature profiles) this process by identifying, as well as by distinguishing, specific features used by people in their decision-making process.

Based on the literature review, some analysis of author verification was conducted to investigate traditional features without ranking and or attempting to understand those features, in order to identify authors on only one platform. Generally, the analysis was positive and provides empirical evidence that shows that lexical features are the most powerful feature for unified user verification on the four modern platforms (Twitter, Facebook, Email and Text messages) and should be examined over multi-platforms to determine the level of feature unification.

Having said that, across all four platforms, a performance of 9.46% was achieved. In addition, this work has technically improved upon using isolated individual platforms. The results seem to suggest that this approach is more effective and warrants further investigation. Furthermore, this approach has not been suggested or used before across most modern platforms (Text message, Twitter, Facebook and Email) together. Although these results need to be further researched, they are positive, and suggest that it is possible to reach even better performance.

More importantly, a minimum amount of data is required to achieve reasonable performance. In addition, if a system that uses four platforms is introduced, obtaining data from those four platforms should ensure meeting the minimum requirements, although it is more difficult than using one platform. However, and more importantly, from a pragmatic perspective, using a single classifier approach is pragmatically better than using an individual approach.

Moreover, the analysis of unified features for platforms provides a continuation of the previous evidence on independent platforms that lexical features are the most powerful feature for unified user verification on the four modern platforms together (Twitter, Facebook, Email and Text messages), and should be examined over multi-platforms to determine the level of feature unification.

## 5.4 Portability Feature-Based User Verification Approach

This section will explore the details of the consecutive experiments that have been undertaken to investigate the proposed common features across messaging systems. In the previous experiment on authorship verification, it was assumed that the training and test data were drawn from the same distribution to match profiles, and the lexical features were shown to be a very powerful, especially under conditions where training and test documents come from the same thematic areas, and to become common across platforms, but for this section, this assumption is different. This is due to domain mismatched profiles across platforms, for example Facebook posts versus Twitter tweets (portability). These experiments will focus on providing the empirical basis for whether this approach would work, initially through exploring the portability of some specific characteristics or features that are portable across platforms, to draw conclusions about its authorship and to understand the variability and difficulties in successfully identifying individuals. The fundamental challenge is how to find the general patterns and bridge them across heterogeneous samples of different platforms for author verification.

This section aims to identify the user through similarities or the matching of common user features on different platforms by verifying user feature vectors against multi-platforms. It will also investigate whether lexical features are common or not, as these were found in the previous experiments.

The portability profiling approach is based on the portability of the individual most discriminating features that the profiling template generates from user platforms. It works by verifying the users' most discriminatory features and then comparing these features to other platforms.

A set of experiments were conducted with different settings to investigate the effectiveness of different features for verifying a given author's different samples.

### 5.4.1 Methodology for the Portability Feature-Based User Verification Approach

The methodology for the portability feature-based user approach was implemented as follows:

- The dataset which contains all users' trained samples (i.e. Twitter, Facebook, SMS text message, and Email samples) were put through a process to extract the features of each user in order to prioritise them in terms of discriminative information. Therefore, the RF deals with this as a two-class classification problem, as shown shown in Figure 5-2.

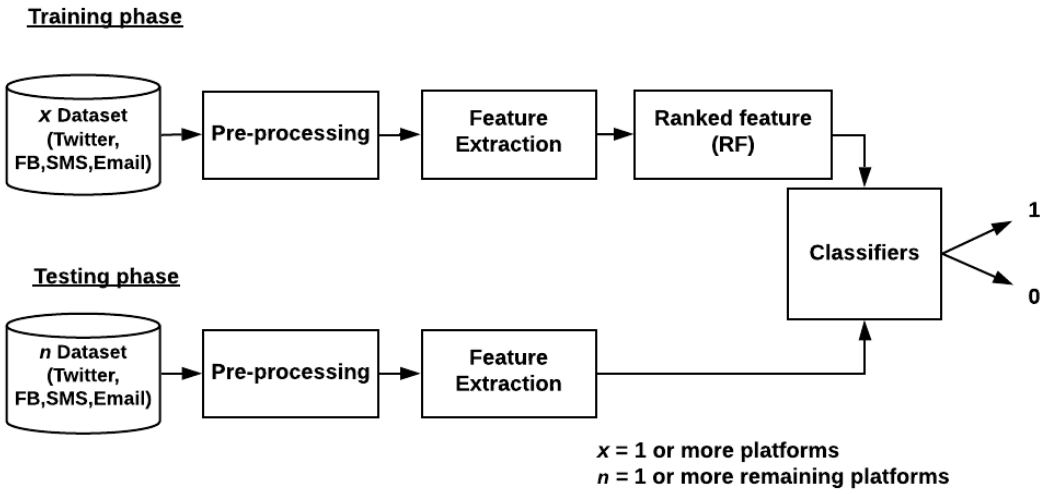



Figure 5-2: Methodology for the portable feature-based user verification approach

- To create user ranked features, the Random Forest algorithm measure was utilised to rank input user features according to their discriminative capability, and these were used in this experiment in the training stage.
- Three different classification algorithms were used to find the optimum classifier for verifying a given message's authorship. This includes support vector machine (SVM), Random Forest (RF) and Gradient Boosting (GB). Each classifier was tested using a different set of features and different train/test split ratios, as presented in the next section. Also, equal error rate (EER) has been used to evaluate the performance of the classification algorithms (Jain et al., 2007).
- In the classification procedure for the training phase, all data from authorised users was utilised, and the size of the imposter's data was reduced to be similar to the size of the authorised users' data; while in the testing phase, all data from authorised and imposter users were utilised.

#### 5.4.1.1 Expand Methodology

In order to conduct portability across platforms for identifying common features, two types of investigations have been undertaken as follows: Testing platforms_vs_platforms, including all 275 features, and testing different types of stylometry features in order to investigate the impact of stylometric features. Table 5-16 below shows the experimental combination.

Table 5-16: Experimental combination

| Testing platforms_vs_platforms | Testing different types of stylometry features |
|---|---|
| 1- Single platform_vs_Multipleplatforms<br>2- Two platforms_vs_ Two platforms<br>3- Multiplatforms_vs_Single platform | 1- Lexical features<br>2-Syntactic features<br>3- Structure features<br>4- Social Network Specific features<br>5- Emotional features |

The desired results were reached with the help of the historical corpora available. For all four corpus samples, each classifier was tested in a setting of "70/30" since this setting was revealed to be the best setting for population, user-based and unified techniques from among other settings, each of which was repeated for the number of authors in the dataset by using a one-vs-all approach. This resulted in the total number of tests shown in Table 5-17 below:

Table 5-17: Total number of experiments for all datasets

| Platform | #Users | #Classifiers | Configuration | Total experiments |
|---|---|---|---|---|
| **Twitter** | 41 | 3 | 1 | 123 |
| **Text Message** | 26 | 3 | 1 | 78 |
| **Facebook** | 46 | 3 | 1 | 138 |
| **Email** | 47 | 3 | 1 | 141 |
| | | | ***Total** | 480*2= 960 tests |

### 5.4.2 Experimental Results

*Testing platforms_vs_platforms which contains the following:*

1- Investigating single platform_vs_multipleplatform results

To identify common and portable features (Single platforms_vs_Single platforms), four experiments have been conducted as follows: Train Twitter_vs_Facebook, SMS, and Email results; Train SMS_ vs_ Twitter, Facebook, and Email results; Train Facebook_vs_ Twitter, SMS, and Email results; Train Email_vs_ Twitter, SMS and Facebook results.

- Train Twitter_ vs_ Facebook, SMS, Email platforms

Table 5-18: Portability Single platform_vs_Multipleplatform results

| Test ID | Feature tested | Performance EER (%) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | SMS | | | FB | | | Email | | |
| | | SVM | GB | RF | SVM | GB | RF | SVM | GB | RF |
| Test 1 | Top 10 | 58.31 | 47.64 | 47.75 | 44.65 | 45.63 | 46.07 | 46.91 | 45.53 | 46.02 |
| Test 2 | Top 20 | 58.76 | 46.73 | 48.84 | 44.97 | 45.01 | 45.57 | 46.62 | 44.66 | 45.86 |
| Test 3 | Top 30 | 56.3 | 47.31 | 48.18 | 44.36 | 45.12 | 45.61 | 46.17 | 46.22 | 46.68 |
| Test 4 | Top 50 | 53.47 | 48.01 | 48.61 | 44.36 | 46.09 | 45.46 | 46.64 | 46.6 | 47.26 |
| Test 5 | Top 100 | 53.78 | 47.6 | 48.28 | 43.86 | 45.35 | 46.86 | 47.82 | 46.2 | 50.85 |
| Test 6 | All | 51.97 | 47.8 | 50.29 | 44.9 | 45.7 | 45.31 | 48.55 | 45.66 | 48.42 |

An overview of the above table shows that the performance is poor for all platforms. The best results for these poor performances went to Twitter_ vs Facebook, which achieved an EER of 43.86%, and so it does not seem to work as expected on these platforms, and it was not clear that there was a linguistic commonality between the Twitter_ vs_ Facebook, SMS, Email platforms.

- Train SMS_ vs_ Twitter, Facebook, and Email platforms

| Test ID | Feature tested | Performance EER (%) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Twitter | | | FB | | | Email | | |
| | | SVM | GB | RF | SVM | GB | RF | SVM | GB | RF |
| Test 1 | Top 10 | 47.75 | 47.79 | 47.09 | 48.12 | 49.55 | 48.2 | 48.29 | 48.12 | 47.79 |
| Test 2 | Top 20 | 48.67 | 49.43 | 49.64 | 48.8 | 50.26 | 47.31 | 49.15 | 47.76 | 47.11 |
| Test 3 | Top 30 | 48.17 | 48.41 | 49.39 | 47.93 | 49.5 | 46.85 | 48.13 | 48.3 | 49.48 |
| Test 4 | Top 50 | 48.07 | 46.75 | 46.81 | 48.42 | 49.85 | 47.68 | 47.2 | 47.56 | 48.15 |
| Test 5 | Top 100 | 47.76 | 48.34 | 48.58 | 48.16 | 50.31 | 44.85 | 48.94 | 47.98 | 45.12 |
| Test 6 | All | 50.19 | 47.95 | 48.43 | 49.15 | 50.13 | 45.19 | 50.61 | 47.9 | 48.44 |

The above table shows that the performances is poor for all platforms. The best result for these poor performances went to SMS_ vs Facebook, which achieved an EER of 44.85%, but it does not seem to work as expected on these platforms as it is not clear that there was a linguistic commonality between the SMS_ vs_ Twitter, Facebook, and Email platforms.

- Train Facebook_vs_ Twitter, SMS and Email platforms

| Test ID | Feature tested | Performance EER (%) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Twitter | | | SMS | | | Email | | |
| | | SVM | GB | RF | SVM | GB | RF | SVM | GB | RF |
| Test 1 | Top 10 | 48.69 | 45.69 | 46.31 | 50.86 | 46.51 | 47.41 | 46.11 | 44.12 | 46.51 |
| Test 2 | Top 20 | 47.97 | 45.82 | 45.69 | 50.94 | 46.04 | 46.54 | 46.47 | 44.11 | 47.56 |
| Test 3 | Top 30 | 47.8 | 45.21 | 45.52 | 51.91 | 45.78 | 47.71 | 44.14 | 43.53 | 46.99 |
| Test 4 | Top 50 | 46.86 | 45.71 | 46.48 | 53.07 | 46.22 | 48.73 | 46.12 | 43.48 | 47.43 |
| Test 5 | Top 100 | 45.99 | 45.29 | 46.82 | 50.19 | 46.52 | 47.67 | 49.4 | 43.86 | 48.37 |
| Test 6 | All | 47.56 | 45.5 | 46.38 | 46.71 | 46.29 | 47.14 | 50 | 42.9 | 47.44 |

An overview of the above table shows that the performance was poor for all platforms. The best result for these poor performances went to Facebook_ vs Email, which achieved an EER of 42.9%, and it does not seem to work as

expected on these platforms, and it is not clear whether there was a linguistic commonality between the Facebook_vs_ Twitter, SMS and Email platforms.

- Train Email_vs_ Twitter, SMS and Facebook platforms

| Test ID | Feature tested | Performance EER (%) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Twitter | | | SMS | | | Facebook | | |
| | | SVM | GB | RF | SVM | GB | RF | SVM | GB | RF |
| Test 1 | Top 10 | 45.02 | 47.35 | 46.23 | 49.02 | 49.41 | 50.07 | 47.02 | 49.03 | 50.28 |
| Test 2 | Top 20 | 43.97 | 46.3 | 47.38 | 51.55 | 49.85 | 48.71 | 46.6 | 48.97 | 47.88 |
| Test 3 | Top 30 | 45.55 | 45.43 | 46.99 | 48.81 | 49.96 | 49.89 | 47.85 | 48.79 | 48.82 |
| Test 4 | Top 50 | 46.11 | 45.59 | 45.53 | 47.06 | 50.13 | 49.28 | 47.52 | 49 | 49.5 |
| Test 5 | Top 100 | 47.79 | 44.79 | 47.88 | 52.38 | 49.63 | 48.14 | 46.95 | 48.7 | 49.81 |
| Test 6 | All | 49.52 | 45.15 | 47.16 | 48.61 | 49.43 | 49.51 | 46.88 | 49.16 | 49.02 |

The above tables show that the performance was poor for all platforms. The best results for these poor performances went to Email_ vs Twitter, which achieved an EER of 43.97%, and it does not seem to work as expected on these platforms and it is not clear whether there was a linguistic commonality between the Email_vs_ Twitter, SMS and Facebook platforms. Table 5-19 below shows the user performance - best and worst for single platform tests.

Table 5-19: Best and worst users in portability single platform tests

| Test description | | Best | | Worst | |
|---|---|---|---|---|---|
| Train | Test | User ID | EER% | User ID | EER% |
| Twitter | FB | 17 | 30.5 | 4 | 55.1 |
| Twitter | T | 12 | 26.9 | 2 | 60.0 |
| Twitter | E | 16 | 12.5 | 17 | 47.6 |
| T | Twitter | 12 | 15.3 | 8 | 56.7 |
| T | FB | 15 | 37.5 | 4 | 57.1 |
| T | E | 3 | 31.8 | 15 | 64.2 |
| FB | Twitter | 23 | 28.19 | 39 | 55.6 |
| FB | T | 3 | 12.5 | 15 | 57.1 |
| FB | E | 20 | 38.8 | 24 | 62.7 |
| E | Twitter | 6 | 42.0 | 18 | 60.8 |
| E | T | 12 | 44.3 | 13 | 57.6 |
| E | FB | 19 | 37.0 | 40 | 55.4 |

The results for Single platforms_vs_Single platform show that the performance results of all platforms are poor, and unfortunately it does not seem to work as expected on these platforms. It is not clear whether there was a linguistic

commonality between the platforms for single platforms_vs_single platform, therefore two platforms_vs_ two platform has been investigated, as described in the next section.

*2- Investigating two platforms_vs_ two platform results*

Further experiments have been conducted in an attempt to identify common and portable features (two platforms_vs_two platforms (train/test)). Seven experiments were conducted as follows: Train Twitter and Facebook_ vs_ SMS and Email results, Train Twitter and SMS_ vs_ Facebook, Email results, Train Twitter and Email_ vs_ Facebook, SMS results, Train Facebook and SMS_ vs_ Twitter, Email results, Train Facebook and Email_ vs_ Twitter, SMS results, Train SMS and Email_vs_Twitter, Facebook results, Train Email and Twitter _vs_SMS, Facebook. In addition, the GB classifier was used and was tested in a setting of "70/30" since this setting and classifier were found to be the best setting and classifier in the previous experiments on population, user-based and unified techniques from among other settings.

Table 5-20: Portability Two platform_vs_Two platforms results

| Train | | Test | | Performance EER (%) |
|---|---|---|---|---|
| Twitter | FB | SMS | Email | 42.89 |
| Twitter | SMS | FB | Email | 41.89 |
| Twitter | Email | FB | SMS | 41.08 |
| FB | SMS | Twitter | Email | 42.15 |
| FB | Email | Twitter | SMS | 43.76 |
| SMS | Email | Twitter | FB | 44.25 |
| Email | Twitter | SMS | FB | 41.08 |

The overview in Table 5-20 above shows that the performance was poor for all platforms. The best results for these poor performances went to Twitter and Email_ vs_ Facebook, SMS and Email and Twitter _vs_SMS, Facebook, and both

achieved an EER of 41.08%. Table 5-21 below shows the user performance - best and worst - for two platform tests.

Table 5-21: Best and worst users in Two platforms vs Two platforms tests

| Test description | | Best | | Worst | |
|---|---|---|---|---|---|
| Train | Test | User ID | EER% | User ID | EER% |
| Twitter, FB | T, E | 12 | 31.4 | 1 | 53.0 |
| Twitter,T | FB,E | 16 | 37.2 | 6 | 53.0 |
| Twitter,E | FB,T | 16 | 36.6 | 11 | 52.0 |
| FB,T | Twitter,E | 18 | 30.1 | 15 | 51.4 |
| FB,E | Twitter,T | 12 | 32.2 | 11 | 51.2 |
| T,E | Twitter,FB | 12 | 31.3 | 3 | 57.1 |
| E,Twitter | T,FB | 16 | 36.6 | 17 | 51.0 |

The results for Two platforms_vs_Two platforms show that the performance results for all platforms are poor, and unfortunately it does not seem to work as expected on these platforms, and it was not clear whether there was linguistic commonality across two platforms_vs_two platform. Therefore two multiplatforms_vs_ single platform has been investigated and is discussed in the next section.

*3- Investigating multiplatform_vs_single platform results*

Further experiments have been conducted in an attempt to identify common and portable features (Multiplatform_vs_Single platform results (train/test)). Four experiments were conducted as follows: Train Multiplatform_vs_Twitter results, Train Multiplatform_vs_Facebook results, Train Multiplatform_vs_SMS, and Train Multiplatform_vs_Email. In addition, the GB classifier was used and was tested in a setting of “70/30” since this setting and classifier were found to be the best setting and classifier in the previous experiments on population, user-based and unified techniques from among other settings.

Table 5-22: Portability multiplatforms_single platforms results

| Train | Test | Performance EER (%) |
|---|---|---|
| Multiplatform | Twitter | 40.76 |
| Multiplatform | Facebook | 43.42 |
| Multiplatform | SMS | 45.56 |
| Multiplatform | Email | 42.32 |

An overview of Table 5-22 above shows that the performance was poor for all platforms. The best result for these poor performances went to Multiplatform_vs_Twitter which achieved an EER of 40.76%. Table 5-23 below shows the user performance - best and worst - of two platform tests.

Table 5-23: Best and worst users in portability multiplatforms_single platforms tests

| Test description | | Best | | Worst | |
|---|---|---|---|---|---|
| Train | Test | User ID | EER% | User ID | EER% |
| Multiplatform | Twitter | 12 | 20.2 | 13 | 54.5 |
| Multiplatform | Facebook | 3 | 39.9 | 2 | 56.5 |
| Multiplatform | SMS | 5 | 30.9 | 3 | 75.5 |
| Multiplatform | Email | 18 | 34.0 | 4 | 53.5 |

The approach does not seem to work as expected on these platforms and it is not clear whether there was linguistic commonality across multiplatforms_vs_ single platforms. Therefore, in order to test certain linguistic characteristics, the most influential features that have been powerful on single platforms such as lexical and syntactic and so on, have been investigated and are discussed next.

**B- Testing different types of stylometry features**

Further experiments have been conducted on whether the common and portable features can be identified for testing different types of stylometric features, for example lexical, syntactic, structure, specific short messages, and emotional features. The methodology used is the same as the methodology for the population-based platforms versus platform authorship verification; however, in this experiment some specific features have been examined such as lexical features on Facebook (F1-F50) versus lexical features on Twitter (F1-F50), and

the GB classifier has been used and was tested in a setting of “70/30”, since this setting and classifier was revealed to be best in the previous experiments for population, user-based and unified techniques from among other settings. The goal was to investigate the impact of stylometric feature types, which includes the following:

Table 5-24: Results for different types of stylometric features using population feature selection

| Test ID | Features tested | | Description | Performance EER (%) |
|---|---|---|---|---|
| Test 1 | Char based (F1-50) | Lexical | Character-based features (features 1- 50), which count the frequency of specific characters were tested. | 40 |
| Test 2 | Punctuations (F51-58) | Syntactic | A set of punctuations listed from features 51-58 were tested. | 41.31 |
| Test 3 | Function words (F59-208) | Syntactic | A set of function words listed from features 59-208. | 41.47 |
| Test 4 | No of sentences (F209) | Structure | Feature 209, which displays the number of sentences. | 47.12 |
| Test 5 | Word based (F210 227) | Lexical | Word based features (features 210- 227), such as counting the frequency of long words or short words. | 41.60 |
| Test 6 | Short message specific features(F 228-233) | Specific features | A subset of specific features (F228-233) such as frequency of missing words or a period or punctuation in a sentence. | 44 |
| Test 7 | Popular emotional features (F234-275) | Emotional features | A subset of popular emotional features were tested. These emotional features appeared in more than 10% of the data collected from 50 authors. | 49.58 |

The above table shows that the performance of all features tested across platforms was poor. The best results for these poor performances went to Train Multiplatform_vs_SMS for lexical feature (F1- F40), which achieved an EER of 40% for the lexical character-based category (F1-F50). The results for portability for the different specific features across platforms (including single/multiple platforms vs single/multiple platforms), show that the performance across features and across platforms was poor; even in the user-based approach, the

results were poor, and this approach seems it does not work on these specific platforms, and it is not clear whether there was linguistic commonality across platforms.

#### 5.4.2.1 Further Analysis of Portability

As pointed out earlier, no previous studies have focused on four platforms, however, some researchers have studied two corpora such as Twitter and Facebook. Further investigation has been conducted to explore the reasons for these features not being portable between platforms. The point is to review some of the evidence from previous studies as much as possible for different platforms as part of the investigation. Lichtenwalter et al. (2010) and Backstrom et al. (2011), studied two social networks, although their problem is essentially different from the problem of this current research, and they used different platforms. They state that the major challenges were from the sparsity of real social networks, and the very small fraction of potential links in the network due to the strong disproportion in writing styles that users have the potential to form on different platforms.

Among other studies, Mikros (2007) attempted to investigate topic influence on authorship attribution by using two corpora with different techniques. They state that the other major problem is from topics that correlate with authors in many available text corpora, and they state that many stylometric variables actually discriminate a topic rather than just the author. However, forensic, intelligence and security applications seek to identify authors regardless of topic (Madigan, 2005).

While a significant body of research has been conducted into homogeneous social networks, there has been no work on capturing the general principles across heterogeneous messaging systems. The questions that arise are: What

are the core mechanisms by which features evolve in different messaging systems? To what extent can common feature vectors and patterns of users be identified and then be portable between platforms? These questions reveal the interactive human behaviours that underlie the fundamental patterns of different messaging activities. The solution to this problem could be to create more understanding of human behaviour inside their messaging systems. Further practical investigation has been conducted, yet there is evidence that the features were identified and unified across platforms and the performance of all cases in the population - user and unified classification - were positive, as shown earlier in the Feature Analysis and Unified Sections. In addition, the top 10 common features have been identified, for example *User 1* across platforms, but this may be a common feature and does not mean they have the same value across platforms. To answer this question, a serious of investigations has been suggested in order to investigate all four platforms available per user. The table below shows that *User 1, User 15* and *User 18* have four platforms.

Table 5-25: Top ten discriminating features for users on platforms

| User 1 | | | |
|---|---|---|---|
| Platforms | | | |
| 1 | 2 | 3 | 4 |
| Tw | SMS | FB | Email |
| F55 | F28 | F1 | F1 |
| F214 | F1 | F55 | F224 |
| F53 | F233 | F13 | F213 |
| F215 | F232 | F9 | F52 |
| F1 | F275 | F210 | F49 |
| F40 | F2 | F214 | F53 |
| F2 | F210 | F215 | F229 |
| F3 | F234 | F2 | F44 |
| F58 | F229 | F213 | F43 |
| F227 | F55 | F16 | F40 |

| User 15 | | | |
|---|---|---|---|
| Platforms | | | |
| 1 | 2 | 3 | 4 |
| Tw | SMS | FB | Email |
| F1 | F1 | F1 | F1 |
| F32 | F28 | F213 | F40 |
| F2 | F3 | F224 | F20 |
| F33 | F214 | F275 | F213 |
| F55 | F213 | F9 | F227 |
| F3 | F220 | F22 | F30 |
| F24 | F2 | F214 | F228 |
| F4 | F4 | F3 | F103 |
| F19 | F24 | F2 | F59 |
| F224 | F211 | F219 | F214 |

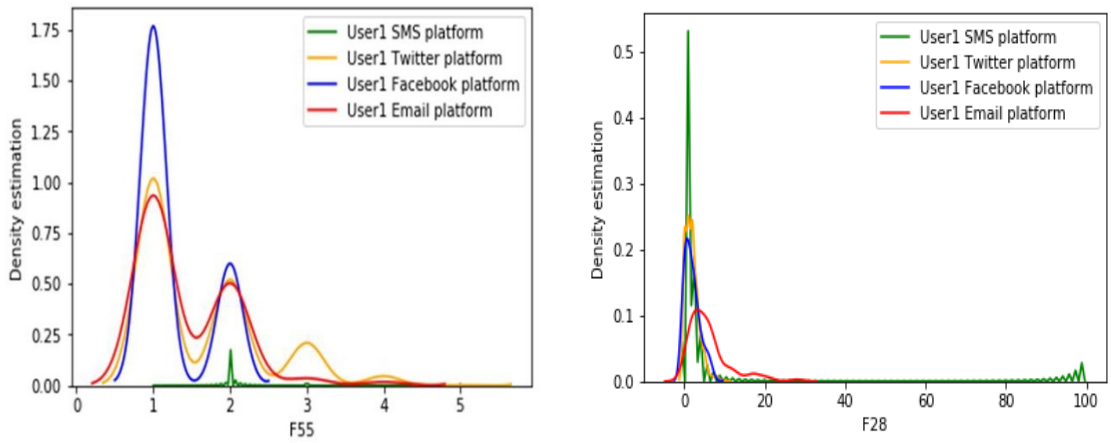
| User 18 | | | |
|---|---|---|---|
| Platforms | | | |
| 1 | 2 | 3 | 4 |
| Tw | SMS | FB | Email |
| F1 | F1 | F1 | F1 |
| F32 | F52 | F233 | F44 |
| F2 | F28 | F55 | F49 |
| F3 | F233 | F4 | F43 |
| F213 | F211 | F229 | F210 |
| F211 | F2 | F213 | F56 |
| F22 | F215 | F215 | F29 |
| F214 | F3 | F218 | F22 |
| F23 | F212 | F214 | F234 |
| F4 | F237 | F216 | F53 |

As shown previously and explained in detail in the previous chapter, Table 5-25 demonstrates examples of the top ten most discriminating features for some users, that is, *User 1*, *User 15*, and *User 18* for four platforms (Twitter, Text message, Facebook and Email). The reason for selecting these users is because

they use four platforms, and the result of their EERs is somewhat not high and not low across the four platforms, and it was necessary to investigate all platforms' feature factors together. As can be seen, there are some common features, for example for *User 1* (i.e., F55, F1, F214 etc.); for *User 15* (i.e., F1,F2, F4 etc.), and *User 18* (i.e., F1,F2, F3 etc). Furthermore, there are also no common features (i.e., F28 for *User 1*; F32 for *User 15*; F52 for *User 18)*.

Further analysis has been conducted in order to demonstrate how these features differ in feature vectors across platforms, visualising feature vectors across platforms (how they look across platforms); by using density estimation, it is possible to see how feature vectors appear across platforms.

Figure 5-3 below demonstrates some examples of some subsets of the most common features for *User 1* across platforms (i.e., F55, F28, F1 and F232) (a full listing of user features for each platform is provided in Appendix H).



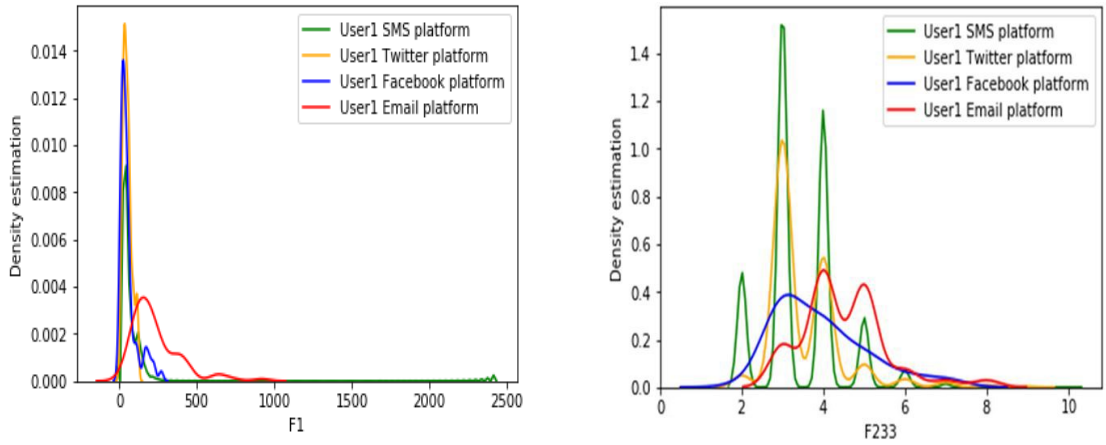



Figure 5-3: Portability: top most common features for User 1

Figure 5-3 shows some of the top discriminating features for *User 1* across platforms, including some features included in the top 10 features across at least two platforms (i.e., F 55, F214), and some features not included across platforms but that just appeared on one platform - not portable (i.e., F28, F233). As can be seen in Figure 5-3 above, feature 55, feature 28 and feature1 seem to be similar, and they seem as if they are identical across the user's platforms, thus they might be portable because the feature vectors are quite similar. On the other hand, although feature 28 was not included in the top features and was not common across platforms, it still appears to be shared across platforms and looks portable; however, the problem is that although these features have the potential commonality to match, they still have different values, and the classifier struggled to pick this up. On the other hand, feature 233, although it was in the top ten, the same as feature 28, clearly looks different and there is no similarity compared to the other features. This indicates that although some features were ranked at the top, whether these features were included in the top common feature or not, there are differences in their values.

On further investigation, despite the fact that they seem portable, their values within the features are different; the similarities between values is a very small

fraction, and also there is sparsity between platforms (Lichtenwalter et al., 2010 and Backstrom et al., 2011). However, an interesting finding may be that the least non-sparsity platforms are SMS Text message and Email as they have some stability in their feature vectors, for example F213 and 214 for *User 15*, as shown in Table 5-25.

One of the possible reasons for why these top features are not portable is that many stylometric variables such as frequently functioning words, commonly used lexical richness measures, and word length are in fact discriminating the topic rather than the author. A study conducted by Mikros et al. (2007) revealed the main impact of author and topic on the dependent variable, and they conclude that that the feature was actually discriminating the topic rather than the author. However, the study was performed on a single platform containing only two authors and on only two topics.

Moving to *User 15*, as was proven earlier, Table 5-25 shows examples of the top ten features that seem to be common for *User 15* ( i.e., F1, F4, F213), although some were not common between platforms, for example F32. As illustrated earlier, although some of these features were common and came top, this does not mean they have the same value and are different enough to be picked up by classifiers, which is why they are not matching. Figure 5-4 shows another example of different subsets of the top most common features for *User 15* across platforms.

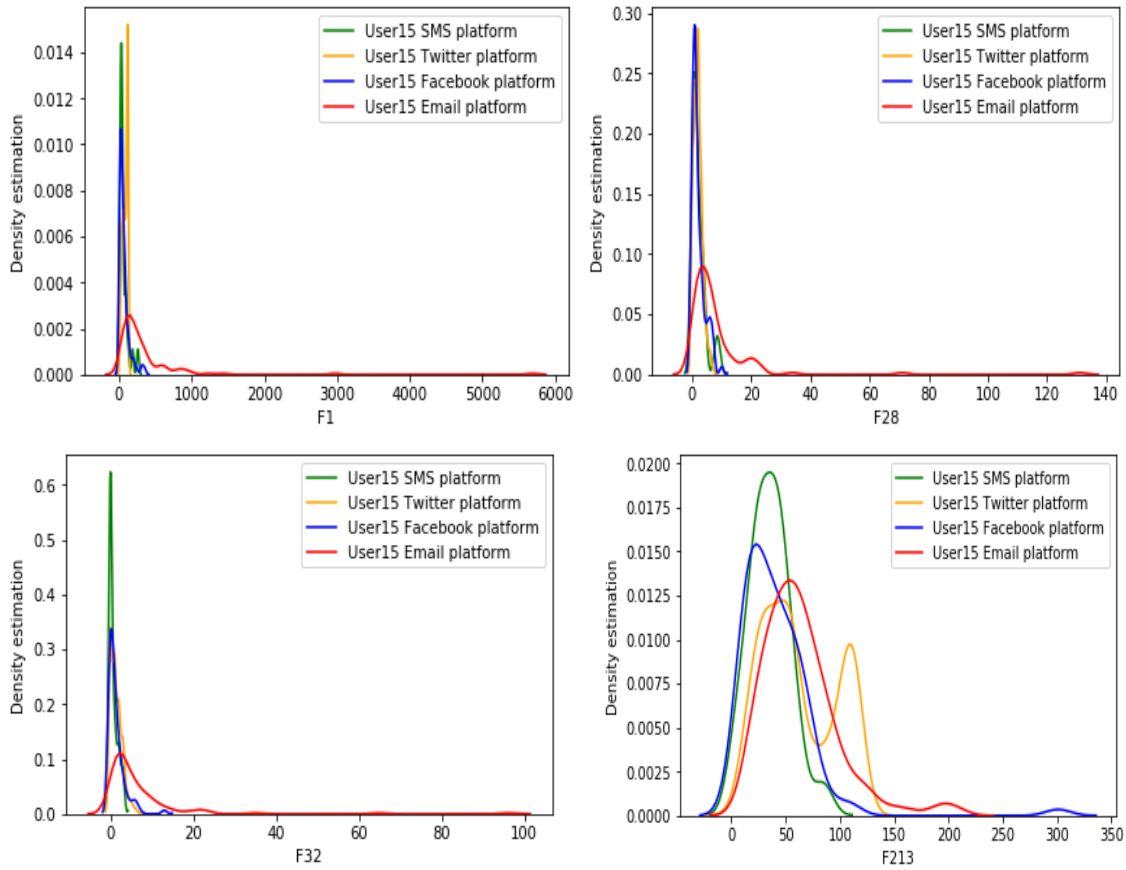



Figure 5-4 : Portability of top common features for User 15

Figure 5-4 shows the top discriminating features for *User* 15 across platforms, including some features that were included in the top 10 features (i.e., F 1, F213), and some features not included in the top ten features across platforms, for example F32; although in both cases, these common feature sets have been identified, as well as the common features for this user. On the other hand, it is still the feature vector that looks like and is similar across platforms and could be portable; while feature 213, although it was clear it cannot be portable to some degree, is clear and looks to have some similarities. All in all, this indicates that although some features were ranked top for users, whether for both cases, these features were included and not included in the top; therefore, there is huge variability, which is why the classifiers struggled to pick up them. As a result of the previously mentioned reasons, some feature vectors have a sparsity and

divergence of feature values, and the features could describe the document itself rather than describing the user (Mikros et al., 2007).

Further investigation has been conducted, and Figure 5-5 below shows some of the top discriminatory features for *User* 18. It can be seen that the feature vectors seem identical across platforms and have common features.

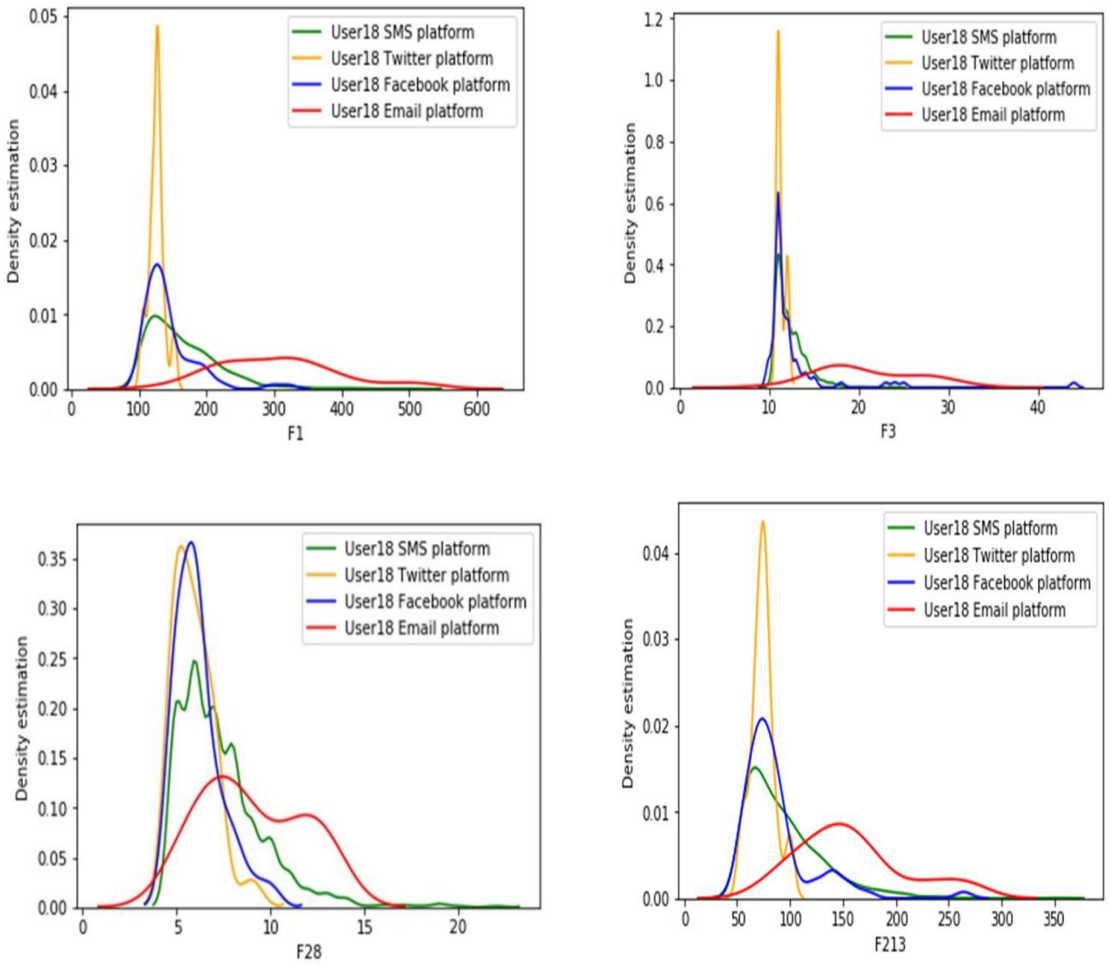



Figure 5-5: Portability top common features for User 18

Figure 5-5 shows the top discriminating features for *User 18* across platforms, including some features included in the top ten features (i.e., F 1, F3, F213), and some features not included in the top ten features across platforms, such as F28. Although in both cases these common feature sets have been identified, as well as the common features for that user, the feature vector still looks similar across platforms and could be portable. While feature 213, although it was clear, is

portable to some degree, it has some similarities. All in all, this indicates that although some features were ranked top for users, whether both cases of these features were included in the top or not, it is clear that there is huge variability between vector values, which is why classifiers struggle to pick them up. As a result of the previously mentioned reasons, some feature vectors are sparse and divergent and, again, the features may describe the document itself rather than describing the user (Mikros et al., 2007).

The main goal of this chapter was to explore the relative performance for common features that can be used across platforms by using the three approaches. The first approach involved feature vector analysis of the most discriminating features for population and user base authors across platforms. The commonalities and differences that exist within the feature set for the population base have been analysed (i.e. the top 10, 20 and feature analysis - common features - between platforms). This includes a comprehensive survey of their interrelationships linguistically with other platforms, for which these subsets of stylometric features would be more reliable in determining authorship from among the population. More importantly, the majority of the most common features on all platforms were lexical. The analysis of the top features (top 10, 20 and 30 features) has been selected because it shows that lexical features have common characteristics when the top ten and twenty characteristics are examined, and to make sure and confirm this, the first thirty were also selected. This showed that lexical is the most common feature across the Twitter, Text message, Facebook and Email platforms, even if the number of features increases to include the top 20 and top 30 features, as shown in Table 5-1, Table 5-2, and Table 5-4.

A feature analysis of the population in the experiments was conducted to address the core research questions which are related to common features among the

population (e.g. features across and between platforms). Therefore, in order to obtain appropriate and sufficient information to create the reference template, the common stylometric features for the top features for multi-platform authorship (Common Feature Vector) among the population have been explored.

A user-based technique has played a major role, and has contributed towards determining that each individual has their own unique writing behaviour and linguistic behaviour features across platforms, for example features 53 and 229 for *Author 1,* as shown in Table 5-7. This was also extended to exploring common features with other different messaging systems. Therefore, further analysis of user-based features was conducted with other different platforms, such as comparing between two, three or four platforms to represent the common feature sets in a more elaborate manner. As shown in In Table 5-8, Table 5-10 and Table 5-11, the authors were selected based on the availability of platforms, for example *Authors 1, 15* and *18* have four platforms; *Authors 21, 25* and *30* have three, and *Authors 48* and *25* have two platforms. When reviewing the pattern of messaging systems usage, it was found that authors had common feature sets between platforms and the lexical feature has been proven to be a very powerful characteristic, as each individual has unique categories within the lexical features. This could help the classifier to identify the user more easily because the results appeared positive when the strongest features were captured, and they differed among authors. This is the first study that has attempted to solve the cross-domain author verification problem by exploiting the most discriminating features with cross-domain datasets. With respect to the research question regarding exploring the feature vector of what commonalities and differences exist within the feature set across the platforms, the exploration of the feature

vector has been analysed for cross-platform authorship (Common Feature Vector for User-Based).

The second approach unified the most discriminating features for authors across platforms and focused on understanding how an author profile can be unified across platforms of historical corpora. The main goal was to show, empirically, the unified user profiles across platforms, and also to show an understanding of the impact of unifying the most discriminating features for users on multi-platforms. This is the first study to solve the cross-platform author verification problem by exploiting certain features across modern messaging platforms. This approach has involved bringing together as much data as possible from the profiles that are available across modern platform to give the best result. It has included identifying as much genuine information from users as possible and incorporating it into a classifier; this approach has then suggested what factors can possibly be picked up.

Across all four unified platforms, a performance 9.46% was achieved. In addition, this work has technically improved upon using isolated individual platforms, as the results seem to suggest that this approach is better. Furthermore, this approach has not been suggested or used before. Although these results need to be further researched, they are positive, and suggest that it is possible to reach even better performance.

The results are indeed positive and yielded an EER of 9.46%, with lexical features showing promising results for unifying users' most discriminating features across platforms (Twitter, Facebook, Email and Text messages). As shown in section 5.3.2, Table 5-13, the best unified features for authors were determined by the top 100 features. Also, it showed that authors had valuable discriminative

information, which could be useful for identifying the most robust features across platforms in different scenarios. For example, the first top 10 of the most unified discriminative features of *Author 1* differs from *Author 15* and from *Author 18*, which strongly supports that there are some categories of lexical features that are robust and unified for authors, and also there are some robust features of authors that differ independently across platforms. This should provide a way for common lexical features to be made portable as a very powerful feature across platforms. However, for some authors, their unified feature usage was fairly poor regarding being unified because their features are not robust enough across platforms. More importantly, a minimum amount of data is required to achieve reasonable performance. In addition, if a system that uses four platforms is introduced, obtaining data from those four platforms should ensure meeting the minimum requirements, although it is more difficult than using one platform. However, and more importantly, from a pragmatic perspective, using a single classifier approach is pragmatically better than using an individual approach. Therefore, a framework for unified features needs to be developed to determine the approximate possible behaviour of each user, and the features of users who successfully show unified behavioural profiling.

The third approach in this section focused on portability features, which is a novel approach, as it is the first study that has explored and attempted to solve the portable cross-domain author verification problem by exploiting the most discriminating features across population and user-bases for modern platforms (Text message, Twitter, Facebook and Email together) to find the most discriminating features across-platforms; and finally, matching features across platforms by trying as many different methods as possible to make them portable

(to the best of my knowledge) and for verification; therefore, it has not been possible to conduct and carry out a comparison with other research.

In this PhD, the proposed approach has included different techniques and different experiments, and various procedures have been attempted with the help of the historical datasets available to investigate the portability of common features across platforms. In addition, efforts have been made to exploit the similarities of common types of feature vectors for author verification to find a possible technique for matching with a high degree of confidence. As shown in the experimental results (in section 5.3.2), there are two main types of investigations that have been undertaken: firstly, testing platforms_vs_platforms (including testing four platforms against another four platforms; one platform vs one platform; two platforms vs two platforms, and three platforms vs one platform). For all four platforms, they were compared against each other to examine the common features to be used for portability. The second investigation included testing the potential of different types of stylometric features with the most effective categories for cross-four domain author verification.

However, through the experiences of all these possible attempts of portability, the performance for the portability of stylometric features across platforms to build the author's profile template can be said to be poor and did not achieve the desired results for verifying authors. It can be said that the findings for unified are better than for portability. In the portability section it can be said that the possibility of portability is that that lexical features are closer to the thematic area, and thus provide an effective author discriminator across these platforms. The best performance finding was an EER of 40% across platforms for lexical character-based features, as shown in Table 5-24. Moreover, syntactic features achieved an EER of 41.47%. This explains that lexical features are the closest features for

user portability that may be transferred with the author across platforms. When the features of these platforms were examined, it has been explored that lexical features are perhaps the closest match in similarities (portability) at the end of the spectrum - it may be common in these platforms and this is perhaps not because of the size but rather because of the composition of the message itself. For example, lexical features seem to play a larger role than other features across platforms. Lexical features are the most common feature for users who have more than two platform, because they are involved in more than one platform. Regardless of the size of the message or the nature of the platform, because in the end it seems there are some words that suggested and refered to this user, and the closest feature that indicates to that user is the lexical features.

In addition, the performance for portability classification was not positive, but in reality, the value of feature vectors across platforms showed divergent behaviour and sparsity similar to some social messaging systems (Lichtenwalter et al., (2010) and Backstrom et al., (2011)). Even on private platforms such as SMS Text message and Email, they showed huge variability because Text messages typically involve short text messages while Email can be used for formal and official correspondence (Chin et al., 2014). This is why classifiers struggle to pick up features and improve performance, because most of the common features, although they exist, still have different values. This investigation showed that the lexical features maybe closer and may provide some common features that can be transportable with authors. They are the most powerful features because they achieved the best result compared to all the results for portability, and within the category of lexical features, character based is the closest feature and is the most effective category across platforms, as shown in Table 5-24.

The key challenge of common features comes from the sparsity of features on platforms due to the strong disproportion in writing styles between platforms, although they have the potential to form matches. The majority of prior work on authorship attribution has focused on the traditional authorship problem that deals with single-domain datasets, and little research has focused on across platforms. To the best of my knowledge, there has been no previous research that has collected data on the most important modern platforms currently available in real life, such as Email and Text message, as well as means of social communication, and using the corpora of the English language to conduct the comparison.

Although lexical features have been proven to be a reliable author discriminator feature in many studies for single platforms, most other studies' techniques for lexical features have focused on the sub-word level and used a specific platform, as it has been assumed that it is very difficult to trace conscious linguistic usage across platforms. Other variables have been used in an attempt to capture the vocabulary size used in a text. These measures should also be topic independent for a specific platform and not for multi-platforms, and since vocabulary richness is an author characteristic, it should not correlate with topics on different platforms. Backstrom et al., (2011) state that "a major challenge of the link feature problem which results from the sparsity of real social networks, which mean that the existing links between nodes are only a very small fraction of all potential links in the network".

Therefore, the portability approach may not be appropriate and so it was advisable not to continue with the same methods, platforms or procedures, or the same number of samples and the same number of users, to find robust common features portable across platforms. This is one of the contributions that may be added to the other contributions of this study in that this method may be

unsuitable with the same methods, platforms or procedures, or the same number of samples. The only method that may have been useful is the first method and second on the findings of multi-platform authorship in single-domain as well as unified user features across platforms, as that achieved good performance, with an EER of 9.46%. Across all four unified platforms, a performance 9.46% was achieved. This work has technically improved upon using isolated individual platforms, as the results seem to suggest that this approach is better. Furthermore, it may be more appropriate to understand the nature and unified behaviour of features for authors, as shown in the second method, in order to obtain robust unified features across platforms, because this has the ability to integrate all features on all platforms with each other. Therefore, this integration between platforms and understanding feature unification may lead to identifying more robust common features in addition to the lexical features, which were shown as the best feature for creating a user profile that can be used across platforms, as illustrated in section 5.3.2, Table 5-13.

Further analysis has been conducted, and an analysis of the number of word features has been investigated, since the majority of authors on most modern platforms are affected by this in their writing. Furthermore, it appears among the the top thirty most important features for single platforms (Number of words, F210), as shown in Table 4-4,Table 4-6 and Table 4-8. Therefore, to investigate this feature across platforms together, and to compare its impact, a series of investigations has been carried out. The analysis of feature vector distribution between users across the four platforms together is presented in the next section, which illustrates that the message length feature has some level of discriminative ability.

## 5.5 Message Length Performance

Further analysis has been conducted, and an analysis of the number of word features has been investigated, since this feature appears among the the top thirty most important features for most single modern platforms (Number of words, F210), as shown in Table 4-4,Table 4-6 and Table 4-8.It is imperative to consider the effect and impact of the number of words on the verification process for modern messaging systems. This is because it may be useful for investigators and analysts to know how much confidence there is in an author verification decision, and to what degree this is dependent upon the length of the message. Given the volume of text is often a restricted factor (due to the nature of messaging systems), key to this investigation is a better understanding of what length of message is required to improve performance. For example, a user's Tweets (with a 140 max character length) might be different to the user's Email messages as these can be considerably longer. Threfore, this section investigates the number of words in short messages with regards to what would be required to make a decision. Bearing in mind that previous studies have determined the length of messages based on unrealistic and not genuine messages such as Text messages and Email messages since they are highly private. This will address the research question concerning what length of message is required to provide reliable verification of an author. It includes four types of real messaging systems' samples: Twitter, Facebook, Email and Text message. Following that is the conclusion and a discussion of the findings, along with highlighting the limitations identified in this research.

### 5.5.1 Methodological Approach

The methodology has been divided into two methods: the first method was used to determine the number of words required for each user on each platform for the

four historical datasets, and it has been proposed to base this on the average word and median value for the number of words across the historical dataset. The second method is the verification process.

In the first method, in order to determine and define the number of words for authors required on each platform, the following steps were applied to each author on each platform:

- The average number of words per user on each platform was calculated (a full listing of all calculated average words per user can be found in Appendix I).
- The first median was used to describe the central tendency of the number of word limits for all users' data by calculating the median for each platform; the reason for using the median is to find out the following limits: the lowest number of words, the average number of words, and the longest number of words for that platform. Once the first median was calculated for all users' average number of words, this value is considered to be the longest word length for that platform.
- The research focus is on limited and small words, so the second median was used to calculate the other lowest, and so the longest words have been ignored.

The figures were divided into three groups: the first median was used to determine the longest words, the second median was used to determine the smallest number of words, and the third group in the middle was used to calculate the value between these two (the values between the largest and smallest words).

Table 5-26 below shows the results of statistically splitting the groups of number of words in the experiment based on the average number of words per author on

each platform, for the first and second median (the full statistical calculations of the median of word length limits can be found in Appendix I).

Table 5-26: Number of word groups

| Platforms | Group 1 | Group 2 | Group 3 |
|---|---|---|---|
| Twitter | <10 | 10-13 | 13> |
| Text message | <5 | 5-9 | 9> |
| Facebook | <6 | 6-11 | 11> |
| Email | <25 | 25-60 | 60> |

The second method involved verification procedures. It used the same settings and procedures as in the previous experiment's user based method as follows:

- Splitting data into a ratio of 70/30 for train and test, since it has been shown to be the best from among all other splitings, as discussed in the previous experiments in Chapter Five.
- The Gradient boosting (GB) classifer was used to test the length of word feature, since it is the first time this clasiffer was used with this specifc feature across platforms to advance the state of knowledge and enable a better decision-making process, and it is the best from among all other classifers, as it has been shown in Chapter Five.
- Prioritising the features in terms of discriminative information prior to being applied to a standard supervised training methodology, RF was used for identifying only the most relevant features. The RF algorithm deals with this as a two-class classification problem.
- In train and test, each group was trained and tested based on determining the number of words that were given and specified in the first method.

Figure 5-6 illustrates the process of the methodology, and the experimental approach to the number of words, including the feature for the specified number of words fitted into the classifer in order to class them based on the two-class problem used to verify them, as shown in Figure 5-6.

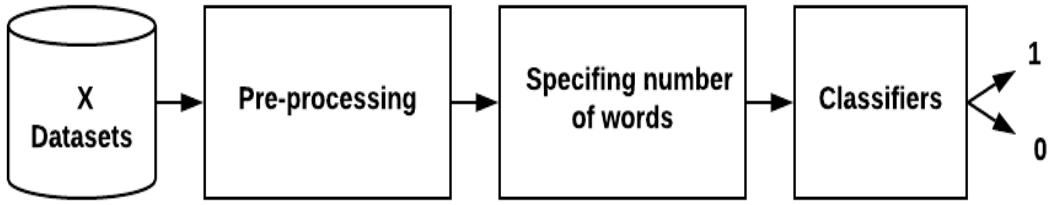
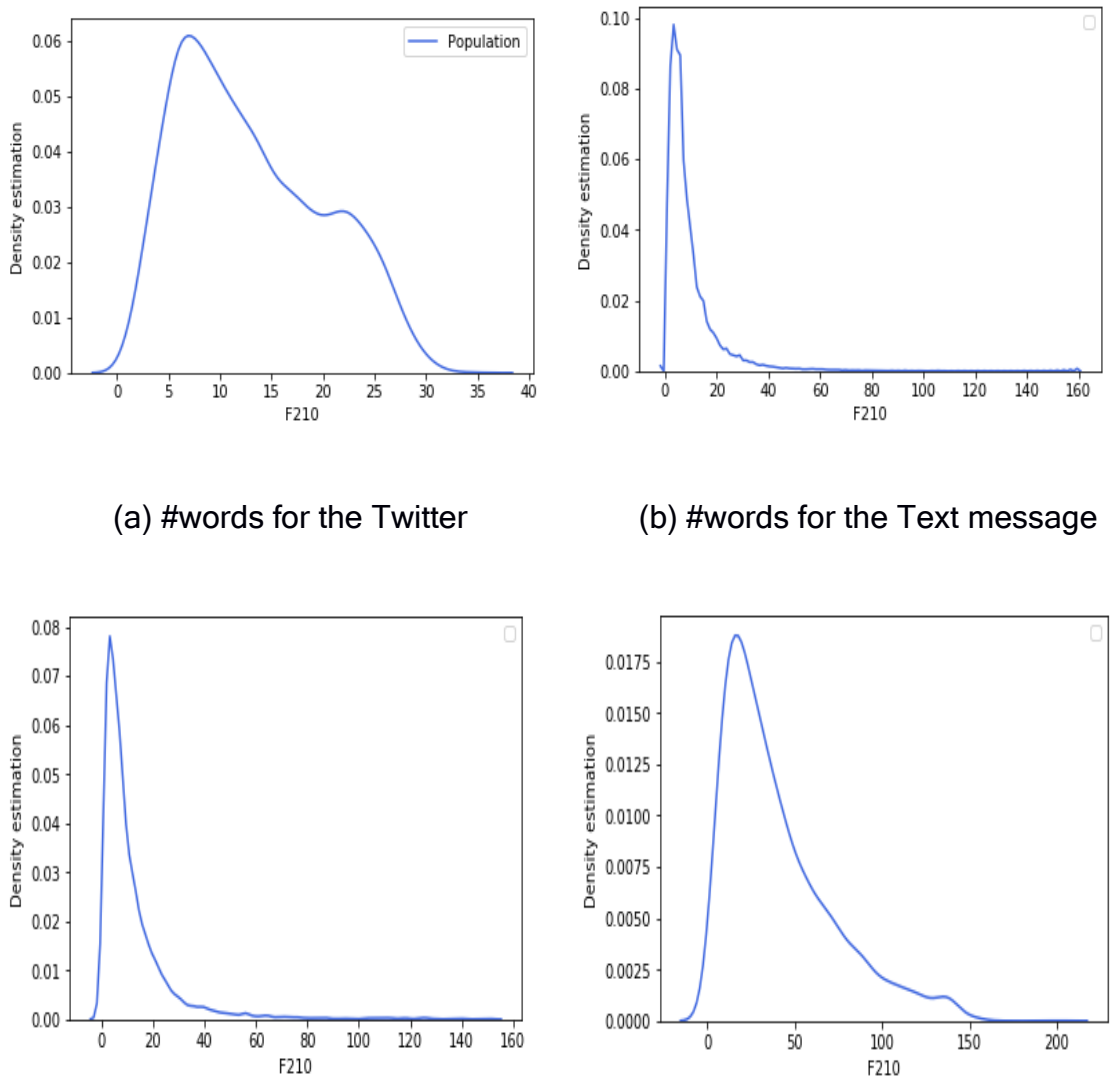



Figure 5-6: Methodology for the number of word-based user verification approach

### 5.5.2 Experimental Results

As illustrated earlier in section 5, visualising the total number of words for the population that is on each platform (Twitter, Text message, Facebook and Email) has been determined in order to understand more about the total number of words for each platform in specific detail, as well as the distribution of word numbers for the population on all platforms in the historical datasets, and the details are presented in Figure 5-7 below.




(a) #words for the Twitter

(b) #words for the Text message

(c) #words the Facebook

(d) #words for the Email

Figure 5-7: Total number of words for population for Twitter, SMS, Facebook and Email platforms

Figure 5-7 shows the total number of words for the population for Twitter, Text message, Facebook and Email. It can be seen that in a comprehensive survey of all platforms, the majority of authors on Twitter tend to use approximately 10 words, while the same thing occurs on the Text message platform, as authors tend to use approximately 10 words, and the same goes for the Facebook platform, as almost the same range of words with a small increase of approximately 10 words tend to be used. However, for Email the situation is different, as the majority of authors of Email tend to use approximately 30 words.

In the case of Twitter, as shown in plot (a), the majority of authors used #words in general that were an average of two to 35 words long in their tweets; however, most authors tend to use a similar number of words on their Twitter account - approximately 10 words. This is expected, as authors have to find a way of being brief and short in their tweet messages using a limited number of words (Sriram et al., 2010).

Similarly, in the case of Text messages, as shown on plot (b), the majority of authors used an average of two to 40 words for their text messages, and most authors tended to use approximately 10 words; again, this is because authors on Text message have to find a way of being concise and short in their messages (Saevanee and Clarke, 2015).

Plot (c) shows that the majority of authors on the Facebook platform used an average of two to 60 words in their posts, and most authors tended to use approximately 10 words; although Facebook is only slightly higher than for Twitter and SMS Text message, which is expected as Facebook messages are usually short in nature (Hussain et al., 2014).

While the majority of authors on the Email platform used words that were an average of two to 150 words long, most of them tend to use between two and 50 words; however, Emails, on the other hand, allow for a large range of flexibility, and they could vary from just a few words to hundreds of words (Li et al., 2015).

Addressing the fundamental research question concerning the relative performance of the information that would be necessary to provide reliable verification of an author, requires measuring and characterising the limitations with respect to message length and composition, to ensure reliable author verification decisions. Dozens of experiments were conducted on the historical

dataset to examine the message length required to understand and enable reliable author verification decisions. Table 5-27 below shows the results for classification performance using the GB classifier. The GB classifier showed better results, as shown in the previous experiments, by splitting the data into 70/30 for train/test, and this was used in this experiment for all groups on all platforms.

Table 5-27: Number of word experimental results

| Platform | Performance EER (%) | | |
|---|---|---|---|
| | Group 1 | Group 2 | Group 3 |
| Twitter | (<10 words) **22.5% (EER)** | (10-13 words) 25.6% (EER) | (13 > words) 23.9%(EER) |
| Text message | (<5 words) 10.6% (EER) | (5-9 words) 10.02% (EER) | (9 > words) **7.9% (EER)** |
| FB | (<6 words) **28.2% (EER)** | (6-11 words) 29.5% (EER) | (11 > words) 31.9% (EER) |
| Email | (<25 words) 15.8% (EER) | (25-60 words) **14.9%(EER)** | (60 > words) 23.3%(EER) |

Table 5-27 demonstrates the performance of the number of words for four platforms: Twitter, Text message, Facebook and Email. Each platform has been categorised into three groups based on the first proposed method in order to investigate what confidence there is in an author verification decision. On average, the best performance of platforms for the experimental results achieved was for Text messages, with an EER of 7.6% if the number of words was more than nine words; followed by Email with an EER of 14.9% if the number of words was between 25 to 60; then, Twitter tweets, with an EER of 22.5% if the number of words was less than ten. Finally, the worst performance from all four platforms and groups was the Facebook platform with an EER of 31.9% if the number of words was more than 11, and the performance of Facebook across groups did not change significantly.

This is expected in terms of the content of the information, as Facebook messages are short in nature (Hussain et al., 2014). Another factor that impacted on performance is that Facebook is used for public purposes, and the author is often writing to various different people on a variety of topics, and so uses a varied number of words, which may make it difficult for classifiers to pick up and verify the author. Unlike the Email platform, which is often directed to a person or to a known group of people, or predefined for who will receive these Emails; thus, Facebook showed poor performance even if the number content of the information was more than 11 words. This shows that if the content information on the Facebook platform is less oriented and accurate, or directed to certain people, the performance for verifying the user on Facebook improves for the above reasons.

In contrast, the best performance was for Text messages, as if the content information and number of words was more than nine words, it achieved good performance at 7.9%. This is expected, because Text messages are sent to specific users and are considered private messages on a personal platform - often one to one - and the individual person's words or writing styles are more familiar for the classifiers. Unlike Twitter or Facebook, the author and texts may be directed to specific people and are not for public use, which suggests that size is less likely to be a determining factor, while the nature of the platform’s use has played a role.

From a different perspective, better results for Text messages, if words (9 >), means that the classifier is also supposed to be better for Email, because it is also based on individual use and the user writes in their own style. It can be noticed that Email achieved a good performance of 14.9% if the number of words was between 25 and 60. This case is similar to Text messages, because it is a private platform and message topics are familiar, so authors use their own writing style and words, making it easier for the classifier to verify the user. From another point

of view, it is noticed that the Text message platform needs more words to provide more reliable performance, while the Email platform needs between 25 and 60 words to ensure reliable performance. This illustrates that the nature of the platforms may also have an effect on the number of words because Text message has a small capacity; therefore, it needs more words to achieve better performance.

Twitter and Facebook messages did not perform better compared to Text messages and Email. This was expected since these platforms (Facebook and Twitter) are similar in nature regarding publicity, which can make it difficult for the classifier to recognise the writing style of the author. On the other hand, it has been noted that Facebook is also worse than Twitter because the capacity of Twitter is as small, and also most authors may be more accurate in their writing and focus more compared to Facebook, as it has a large space for writing. This is another aspect that may contribute toward the better performance of Twitter compared to Facebook.

In general, it can be stated that on the personal and private platforms of historical datasets such as Email and Text message, the increase in the number of words can be more effective for verifying the author's writing style, and the optimal maximum content on the Email platform may be 60 words to deliver good performance. Unlike Twitter and Facebook, the performance improves if the number of words are less, as shown in Table 5-27, so that the classifiers can find any unique number of words that refer to the author to perform well; in addition, since they are public platforms, the topics are diverse, and the writing style is plain as the author is posting to various people. This section has addressed the research question regarding what length of message is required to provide reliable verification of a platform.

### 5.5.3 Investigating User Level Performance

A series of analyses has been conducted, and the authors in this experiment have been selected since they met the previously mentioned conditions. Firstly, they have at least 20 samples across platforms; secondly, they must have four platforms; thirdly, they must have available samples for the number of words feature specified in each group for each platform. In order to investigate the impact and the effect of the number of individual words across the platforms used, and to investigate if it is possible to verify the author based on his/her number of words, Table 5-28 demonstrates the performance of some individual authors across groups and platforms, and how good performance has been selected from all groups. Each author should have at least four platforms combined, the number of samples has to be not less than 20 samples, and the number of words has to be as defined on each platform. Table 5-29 demonstrates the performance of all individual authors across four platforms.

Table 5-28: Some individual classification results by using number of word features

| SMS platform | | |
|---|---|---|
| **User** | **Group** | **Performance EER%** |
| 1 | 1 | 6.6 |
| 1 | 2 | 5.0 |
| 1 | 3 | **4.3** |
| 3 | 1 | 10 |
| 3 | 2 | 8.2 |
| 3 | 3 | **0** |
| 15 | 1 | 12.8 |
| 15 | 2 | 9.4 |
| 15 | 3 | **8.3** |
| **Email platform** | | |
| 1 | 1 | 27.5 |
| 1 | 2 | **23.6** |
| 1 | 3 | 24.2 |
| 3 | 1 | 12.2 |
| 3 | 2 | **10** |
| 3 | 3 | 24.2 |
| 15 | 1 | 36 |
| 15 | 2 | **30** |
| 15 | 3 | 55.8 |
| **Twitter platform** | | |
| 1 | 1 | **29.6** |
| 1 | 2 | 29.9 |
| 1 | 3 | 30.3 |
| 3 | 1 | **22.7** |
| 3 | 2 | 29.7 |
| 3 | 3 | 31.6 |
| 15 | 1 | **20.2** |
| 15 | 2 | 32 |
| 15 | 3 | 20.5 |
| **Facebook platform** | | |
| 1 | 1 | **36.3** |
| 1 | 2 | 44.1 |
| 1 | 3 | 40.5 |
| 3 | 1 | **10** |
| 3 | 2 | 16.2 |
| 3 | 3 | 19.7 |
| 15 | 1 | **28.5** |
| 15 | 2 | 33.4 |
| 15 | 3 | 37.2 |

Table 5-29: All individual classification results for all users with 4 platforms by using number of word features

SMS
# words>9
(group 3), EER
7.9%

| User | EER (%) | Users | EER(%) |
|---|---|---|---|
| 1 | 4.3 | 10 | 9.6 |
| 2 | 2 | 11 | 4.2 |
| 3 | 0 | 12 | 2.6 |
| 4 | 12.6 | 13 | 0 |
| 5 | 5.9 | 14 | 14 |
| 6 | 13.5 | 15 | 8.3 |
| 7 | 9.5 | 16 | 5.7 |
| 8 | 9.6 | 17 | 13.3 |
| 9 | 16.1 | 18 | 5.7 |

Email
#words 25-60
(group 2),EER
14.9%

| User | EER(%) | Users | EER(%) |
|---|---|---|---|
| 1 | 23.6 | 10 | 30 |
| 2 | 9.9 | 11 | 0 |
| 3 | 10 | 12 | 27.5 |
| 4 | 0 | 13 | 16.3 |
| 5 | 13.1 | 14 | - |
| 6 | - | 15 | 30 |
| 7 | 10.6 | 16 | 25 |
| 8 | 37.5 | 17 | 22 |
| 9 | 39.9 | 18 | 12.5 |

Twitter
# words<10
(group 1),EER
22.5%

| User | EER(%) | Users | EER(%) |
|---|---|---|---|
| 1 | 29.6 | 10 | 25 |
| 2 | 0 | 11 | 19.4 |
| 3 | 22.7 | 12 | 8.3 |
| 4 | 39 | 13 | 38.8 |
| 5 | 30.8 | 14 | 32 |
| 6 | 34.3 | 15 | 20.2 |
| 7 | 34.4 | 16 | 19.4 |
| 8 | 21.5 | 17 | 32.6 |
| 9 | 35 | 18 | - |

Facebook
# words<6
(group 1), EER
28.2%

| User | EER(%) | Users | EER(%) |
|---|---|---|---|
| 1 | 36.3 | 10 | 0 |
| 2 | - | 11 | - |
| 3 | 10 | 12 | 35 |
| 4 | - | 13 | 21.7 |
| 5 | 37.1 | 14 | 45 |
| 6 | 29.7 | 15 | 28.5 |
| 7 | - | 16 | 40 |
| 8 | 36.1 | 17 | 40 |
| 9 | 28.3 | 18 | 8.3 |

Table 5-28 demonstrates the performance of some individual authors across groups and platforms, and how good performance has been selected from all groups. Table 5-29 shows the performance of authors using the message length features previously defined for each of the four platforms. From this table, it can be observed that the Text message and the Email platform display better performance compared to Twitter and Facebook. It can also be seen that some users, such as *Authors 1's* EER in Text message was 4.3%; 23.6 for Email; 29.6% for Twitter and 36.3% for Facebook. In this sense, the order of the EER ratio for authors across these platforms is as follows: Text message, Email, Twitter and Facebook, ascending in the sense that the pattern of the author can be determined by the ascending range of relative performance in this order. While some authors,

such as *Author 3* differs, as their EER was 0.0% for Text message; 10.0% for Email; 22.7% for Twitter and 10.0% for Facebook. Furthermore, it can be noted that the difference in the level of the author's pattern according to the relative performance is as follows: Text message, Email, Facebook and Twitter; in this sense, it has been found that Facebook's performance is better than the performance of Twitter for that author, and since Facebook is similar in performance to the Email platform at 10%, this means that the user pattern is closer and exists on these platforms - Text message, Email, Facebook and finally Twitter - in ascending order. While some authors, such as *Author 15* differ, as it can be noted that the pattern can be determined according to this order: Text message, Twitter, Facebook and Email. Therefore, this pattern has addressed and answered the research question regarding what length of message is required to provide reliable verification of an author that are not similar to these platforms' performance for each individual. However, the length of message can provide a reliable verification for some authors across the datasets, as shown in Table 5-29. The ascending order according to relative performance based on the best to the worst performances of the historical datasets, the better performance for these four platforms is as follows: Text message (more than 9 words with an 7.9% (EER)), Email (between 25 to 60 words with an 14.9% (EER)), Twitter (less than10 words with an 22.5% (EER)) and finally Facebook (less than 6 words with an 28.2% (EER)).

## 5.6 Conclusion

In this chapter, multiple investigations have been presented to improve the performance of cross-platform author verification, and a series of investigations have been conducted to explore common stylometric features in both single and cross-domain datasets. The first proposed method is the relative performances

between platforms and the feature analysis of single-domain datasets and, importantly, the majority of the most common features on all platforms of single-domain were found to be lexical for both population and user-based author verification. This provides evidence that these common feature sets could be identified for most authors across the modern four platforms, and the performance for classification was positive, as shown in the previous chapter in Table 4-2 and Table 4-13.

The second proposed method unified users' features across platforms for all unified four platforms. A performance 9.46% was achieved, and this approach has technically improved upon using isolated individual platforms, with the results seeming to suggest that this approach is better. Furthermore, this approach has not been suggested or used before. Although these results need to be further researched, they are positive, and suggest that it is possible to reach even better performance. Importantly, the most optimal classifier for unified experimental studies was GB, which can be used to build successful user unified behaviour profiles within the modern messaging systems. It also provides evidence that these lexical feature sets could be identified for most authors as part of the process, as shown in Table 5-13; in addition, they could be unified across platforms, as shown in Table 5-15. This comprehensive practical study has explored which of the most widely used techniques for author verification are best, and it has shown that lexical features are effective for cross-domain author verification.

In the third proposed method, the ability to solve the cross-domain datasets author verification problem through the portability of discriminating features across platforms was explored. There is a high degree of variability between the linguistic characteristics for platforms such as Twitter, Facebook, Text message and Email, which would suggest that the ability to use information from one platform is not

transferable to another (portability) using the three classifiers used. However, interestingly, an approach that utilises data from multiple platforms in a single classifier does appear to have useful characteristics because the performance on average across the four datasets is under 10%: using data from across the four platforms in a single classifier gives a critical performance, and an advantage is that the volume of training data required for one platform can be reduced in comparison to examining a platform in independent mode.

Based on the current findings for author verification across modern historical corpora, lexical is the most powerful feature for cross modern platform author verification (Twitter, Facebook, Email and Text messages); and within them, the number of characters and number of words (lexical feature), is the most effective category for the historical corpora for author verification. Different features across-domain and single-domain have been analysed and compared in this chapter. This is the first study that has attempted to solve the user features cross-modern platform author verification problem together by exploiting the most discriminating features in this way (to the best of my knowledge).

Further analysis has been conducted in this chapter, and the number of word feature has been investigated to determine the number of words that would be required to ensure the reliable verification of an author across the four modern datasets. This is because it appeared among the the first top thirty important features for most single platforms (Number of words, F210), as shown in Table 4-4, Table 4-6 and Table 4-8. The stylometric feature of length of word improved the performance of an author across platforms based on the optimal word number set has been given for each platform, and this is based on the number of words that are specified for each platform. The findings in this section have determined the best/worst message length in the investigation for each platform by

determining the relative performance and the best and worst word limit for each platform. For example, on average, the optimal length of messages for the experimental results achieved for Text messages was more than nine words, with an EER of 7.9% and the worst if the number of words was less than five, with an EER of 10.6%; the optimal length of Facebook posts was less than six words, as the EER was 28.2.8%; then, Twitter tweets, as if the number of words was less than ten, an EER of 22.5% was achieved. Moreover, the Email message investigation achieved the longest number of words compared to the other corpora, as the optimal number of words was between twenty-five and sixty, and an EER of 14.9% was achieved. The best/worst performance of some authors within each corpus has also been determined (i.e the best author's EER for Email was 0% for *Author 4*, and the worst was *Author 15* with an EER of 30%). The best/worst performance of authors across platforms together has also been determined (i.e. Author 3's performance across platforms was 0%;10%; 22.7% and 10% for Text message, Email, Twitter and Facebook respectively). In addition, it was found that the authors' performances were better across platforms when comparing the results in ascending order according to relative performance for these platforms.

Therefore, this investigation has sought to provide a foundation technique for investigators of length of words on platforms to track the footage of an author, and consider the relative performance based on the limit on words for each platform regarding what is required for reliable verification. It should be borne in mind that a user's writing style has potential to change and is not fixed on most platforms, so the user can change the strategy of writing on one or more of these types of modern platforms. This chapter has discussed a possible solution to the problem of author verification by determining the number of words in all of the historical modern data collected on the four corpora by using two methods: the first method

attempted to determine the number of words required for each author by counting the average number of words across platforms, and by counting the first and second median to describe the appropriate group limits across the four historical datasets. The second method involved performing a verification process based on the relative performance of number of words to determine the word limit for each platform, as well as the performance of individual authors across platforms. In addition, some authors' stylometric features can be used to improve the performance - one of them is the number of words, and this has been discussed and may open the door for investigators to further consider this feature (Number of words, F210), which is a lexical feature across four modern platforms; especially since some of the messaging system platforms differ in determining the number of words used by suspects, for example Facebook and Text message.

To conclude, cross-domain and single-domain author verification of electronic messages may provide a viable solution to problems around forensics and security, and to prevent or repel a variety of direct or indirect criminal activities, such as sending threating or terror-related text messages or spam to gain personal information, groom children, kidnap, murder, or encourage violence. Such approaches are important to protect the international community, especially from messages from terrorist organisations such as al-Qaeda, ISIS and others.

# 6. Chapter Seven: Conclusion and Future Work

This chapter concludes the main achievements of the research and discusses its limitations and obstacles. It also highlights the potential areas for further studies within the security field of author verification of electronic messaging systems.

## 6.1 Achievements of the Research

Overall, the aim of the research was to understand the relative performance and to explore the application of authorship verification to the modern messaging systems of Twitter, Text message, Facebook and Email. The first objective was to explore authorship verification within these individual modern messaging platforms, and the possible common features when using different single-domain datasets for population-based and user-based verification approaches have been found. The goal was to understand and analyse linguistic features, and whether they have a relationship between and among the population and individual authors across the platforms initially set out in Chapter Five. The second aim was to explore unifying with portability, author features across platforms, in order to understand what relationship, if any, might exist within linguistics cross-domain, as set out in Chapter Six. In addition, the investigation has also sought to identify the minimum amount of text required whilst still ensuring a reliable performance for each platform. Chapter Six has presented the application of a series of practical experiments on four novel datasets. Overall, the key achievements of this research are that it has:

1- Identified the main problems in author verification for modern electronic messaging systems, from a security, biometric and forensic perspective. From a security perspective, these messaging systems provide environments for authors to connect with their friends and family. Authors get together in these communication community systems for information

sharing or to build relationships. Authors may assume that messaging systems provide a trusted environment for sharing information with friends and family. Electronic messaging systems service providers need to provide some security mechanisms to verify the authorship of messages or to detect any suspicious messages that do not conform to the writing style of the same author. Additional intelligent security measures have been suggested to ensure the legitimacy of the author. From a biometric perspective, the enrolment process in feature analysis requires the existence of an enrolment sample, which is used to compute the behavioural profile of the user. This sample should contain all possible key combinations in order to effectively recognise the user based on an expected or unexpected set of author inputs. The ability to verify the user does not only have application in the digital forensic dominion, but could also be used as a biometric system modality for use in transparent authentication. From a forensic perspective, authorship attribution is applicable to forensic investigations, and it can be used to determine whether the claimed authorship of a document is valid, and it can also be used to improve spam filtering, or to verify the authorship of threatening Emails by confirming with Facebook messages. Also, it could eventually lead to being applied to fight different forms of cybercrime such as verifying authors of hate speech and defending against paedophiles, grooming children, kidnap, murder, terrorism and violence. However, a greater level of accuracy is required to be suitable for criminal prosecutions, although the research provides hope for the future for certain platforms, as the best experimental results achieved were for Text message, with an EER of 7.97%, and three authors experienced EERs equal to or less than 0.2%; followed by Email with an EER of 12.03%; then, Twitter tweets, with an EER

of 20.28%. Finally, the worst performance from all four platforms was the Facebook platform with an EER of 23.7%. Furthermore, the current research may help to narrow the field of investigation, ultimately reducing the amount of time taken to look for suspects, even though it would not stand up alone as evidence for prosecution.

2- Performed a comprehensive review of the potential usage of techniques for security and forensic purposes by presenting a wide range of techniques for author verification. Author verification approaches play an important role in such cases since it is believed that suspects unconsciously leave stylistic marks in their writing, and therefore it is possible to forensically verify the true author of the text. Thus, lexical features are the most crucial elements in determining the best way of discovering the significant linguistic markers used by an author. Messaging system providers, security experts or forensic investigators can investigate this feature to secure or build user behaviour profiles for their authors who may violate their policy and spread threats, including in relation to terrorism.

3- Explored the requirements that are needed to apply the aforementioned techniques for improving the security of electronic messaging services, ensuring the freedom of society from messages containing hatred and threats by verifying the real author who is using these modern platforms to commit these horrible crimes. It is important to consider the impact of authorship attribution on free speech and the pros and cons of such actions. In the current environment, concerns have arisen among both the public and Internet analysts that the content and tone of certain online interactions, as well as their intent, has evolved to become increasingly negative, and this poses threat to the future of the internet and how it is managed (Rainie et al 2017). Moreover, they state that the internet is being used to promote

extremist causes, which along with ‘fake news’ and ‘foreign trolls’ is having a major effect on public opinion; therefore, authorship attribution is essential on both an individual and international level. In particular, “Anonymity, a key affordance of the early Internet, is an element that many in this canvassing attributed to enabling bad behavior and facilitating uncivil discourse in shared online spaces (Rainie et al., 2017). On the other hand, they explain that such concerns could be used by governments and big businesses to put extra monitoring in place to suppress free speech. Therefore, while authorship attribution is essential to keeping online communities safe and discovering the perpetrators of crimes, it is also important to take into account the potentially negative impacts on free speech. This has been achieved by exploring and evaluating the contribution of the current state-of-the-art technique, that is, individual, population, unified and a portable behaviour profiling technique on different messaging systems whose infrastructure is almost entirely different from each other, including the core modern messaging systems of Twitter, Facebook, Email and Text message.

4- Developed a novel series of experimental studies on author verification for text messages across different messaging systems. Four datasets from real authors’ messages were collected from 50 participants for the core modern messaging systems of Text message and Email, and two popular modern social messaging networks - Twitter and Facebook. It has also involved additional procedures, and conditions have been applied to the data collection; for example, an author had to have at least a minimum of two platforms, and there must have been at least 20 messages available on each corpus for an author. Text message and Email are more challenging than any other messaging system because of ethical reasons as they have very high privacy levels. Even collecting data from Facebook and Twitter is

not a straightforward task, as it is necessary to request the details of a particular user's online social network to obtain the Application Program Interface (API) for that. The post/tweet details are hard to get and can be made available only after convincing the author to share their account details, which in itself is a challenging task and this research offers added worth to researchers in this field by overcoming the obstacle of this challenge, as several types of software were applied to collect samples from authors on different messaging systems. At that time, each of these platforms required a data pre-process to be developed to parse the relevant messaging data, ensuring only relevant data was parsed. Various stylometric features were designed prior to the experiments by using a number of scripts generated to extract features. Next, these authors' extracted feature profiles were prioritised in terms of discriminative information. Then, these profiles were employed and evaluated by using more complex solutions concerning three classification algorithms (SVM, GB and RF). Lexical features showed a very strong common discriminative feature for the individual, population, and unified on most messaging systems together (Twitter, Facebook, Text message and Email) using a multi-class classification problem, and this had a positive impact on the performance of the classifier selected, as it could cover the patterns of most authors across platforms. For the user-based technique, it revealed that each individual has uniqueness in their own linguistic behavioural features, and each user's writing style becomes a biometric signature for that person. Lexical features provide a robust approach for most individuals and yielded good performance. Importantly, the most optimal classifier for both experimental studies was GB, which can be used to build successful user behaviour profiles within the messaging systems, and it was the first time

this classifier has been used for four platforms combined with each other. The performance of this study is encouraging and show the potential for author verification whether in population, user, unified or portable level experiments.

5- Illustrated the most widely used features, which is that this is the first comprehensive study to show empirically and practically which of the most widely used features of author verification are effective for multi-platform author verification whether for single-domain or cross-domain. It has been shown in both cases that lexical features are closer to the thematic area and, therefore, can be used as one of the most effective author discriminators. The study has also explored the sensitivity of widely used features (lexical, syntactic, structure, short message features, and emotion features), and it is the first time these features have been used and tested under cross-platform settings for author verification on Twitter, Facebook, Email and Text message together. Multiple solutions have been presented to explore the performance of multi-platform author verification.

   The relative performance investigation showed that the lexical features are closer and may provide common features, and secondly, the syntactic features; these are the most powerful features, and within the category of lexical features, character based is the closest feature and is the most effective category for lexical cross platforms. Within syntactic features, the punctuation based category is the most effective category for cross-platform author verification, and thus should be effective for author verification processes on most modern messaging systems.

6- Explored the problem of the length of message that would be required to make a decision. This was investigated as the framework of knowledge of

the number of words on most modern platforms is necessary to keep from being negatively exploited, especially for teenagers and other vulnerable groups in society. The results demonstrate the ability to correctly verify users based on their number of words derived from platform dependent and independent features for author verification. A new approach was proposed and used to calculate the number of words required for each author, by counting the average number of words across platforms and counting the first and second median to discover the appropriate limits of the authors' length of words on each and across modern platforms, before performing a verification process based on the limit of words that was statistically stable. To the best of my knowledge, there has been no study or research similar to this current research concept that has collected more corpora from the modern core platforms currently available, in real life, and to explore the reliable length of message required for verification for platforms together. This investigation has shown that the discriminative information of author profiles can be affected by the number of words feature.

## 6.2 Limitations of the Research

Whilst the main objectives of this research have been achieved, a number of limitations associated with the research have been identified. The key limitations of the research can be summarised as follows:

1- There is some limitation in the datasets, although the project collected a certain volume of data, data from more users across more than four platforms would be useful for more extensive evaluation.

2- Some participants like to use two or more platforms, often focusing on one platform or at most two platforms and ignoring the other platforms. This may cause sparsity in the writing styles and language of that author on all

platforms, and this sparsity reduces the chance of exploring more common author features between platforms.

3- Further exploration of portability is required, as the results from exploring the portability of profiles across platforms are somewhat time limited, so the initial results do not provide precise data that could be observed over a long time period.

## 6.3 Suggestions and Scope for Future Work

The main aim of this research was to to explore the application of authorship verification to modern messaging systems through identifying authors' common features across platforms. This is just the beginning of work on author verification across platforms. For further research and exploration, several scenarios could be investigated, which are as follows:

1- The number of participants and data collection could be extended to form a larger dataset. This would enable a better understanding of the nature of common feature changes across platforms, and further evaluation of these changes for unification and transportability profiles in the real world, such as author profile changes across other platforms.

2- One of the future directions could be to evaluate each proposed approach with other problems such as plagiarism detection, document similarity, and applying other language verification such as Arabic, Turkish, Spanish, Italian, and so on.

3- The frequency and time of sending text messages could be added as an additional feature of measurement, as some authors were more active than others, especially on Twitter and Facebook. Authors may range from sending a few texts a day to a few times a month, or even less frequently. In addition, authors send texts at a time that is convenient to them, therefore

when a text message is created at a time that is abnormal to the usual text message pattern of the author, it might be considered suspicious. Also the N-gram technique is one of the possible techniques that will be tested across the four corpora.

4- This research has analysed university students' sample messages across platforms. Each author may have individual preferences for the types of information that they like to share or comment on. For example, sports lovers may create messages related to sporting activities and so on. On the other hand, those interested in politics may create messages that are related to political news, and so on. Analysing authors' topics in their messages based on the person's orientation and interests, could improve the performance and support by finding the most common linguistic characteristics based on their interests or orientation.

5- At the moment, and because of the time constraints, the initial preliminary results suggest that portability does not work so well, so further work needs to focus on exploring this in more detail to look for opportunities where it might work. For one or two users it was more plausible, and further research might create a better understanding of the nature of relationships.

6- Some users may adopt the use of emoticons, abbreviations or any other fashionable behaviours when friends or people from their social messaging circles are using them. If there is no stability in the message writing styles' consistency process, it would be interesting to explore how consistent users' writing styles are over a long period of time. It was assumed in this research that it does not reflect the full picture of real-world scenarios of authors' writing styles if the platform is mainly used for social messaging and social circles communication. However, in reality, authors communicate about a plethora of different topics; therefore, a portability

experiment may not necessarily reflect the variation in data input that could be observed in practice.

## 6.4 The future of Author Verification of Electronic Messaging Systems

The variety and popularity of social network messaging systems such as Text messages, Email, Facebook and Twitter, has seen a rapid increase, with users frequently exchanging messages across a variety of platforms. The rapid spread of the number of messaging platforms has influenced the speed of information sharing and made it more easily accessible, including for extremist groups such as the Daesh terrorist group (IS), Al-Qaeda, Golden Dawn, and others; they provide channels for delivering evil messages straightforwardly and perfectly to the intended party and their target audience. Furthermore, it is popular, especially for those in their 20s, to have multiple messaging systems to enjoy and explore these new platforms, and teenagers are often quick to respond and decide without considering the consequences or where this message comes from, which highlights the potential dangers. In addition, messages may be sent from alleged friends on another messaging platform, and it is possible to discover his or her interests. In amongst the legitimate messages, there can be a host of illegitimate and inappropriate content, with cyber stalking, trolling and computer-assisted crime all taking place. This can even lead to a variety of direct and indirect criminal activities, such as encouraging teenagers to travel to conflict zones (Iraq, Yemen, Afghanistan etc.); grooming children; promoting and facilitating kidnap, murder, terrorism and other forms of violence, as well as sending spam texts to gain personal information. Therefore, it is essential to find a way of tracking users as a response to the illegitimate use of social messaging systems.

Despite many studies having been undertaken to recognise individuals on messaging platforms, this thesis emphasises the need for a robust mechanism for choosing the most important feature sets, whether for a single platform or across platforms, by prioritising the features in terms of discriminative information for user behaviour profiling via finding the most robust features, both population-based (so across all users), and user-based (across the authorised user). This is significant and a fundamental requirement to permit and open the door to investigating the impact of distinctive characteristics on recognition performance, especially for modern electronic messaging systems, as they are diverse and have a variety of uses, with each platform requiring certain writing styles, features and special specifications; such as Email for official use, while Facebook and Twitter are for social networking. The results of this research demonstrate a significant improvement in platform-independent author verification performance for prioritised users (profiles), allowing the forensic investigator to go beyond common features for cross platform optimisation.

Criminals can use all means to reach the target platform and from the methods they use is targeting more than one platform to spread their agenda, often attempting to hide on another platform; however, a substantial number of earlier studies have targeted a single platform, and there has been a lack of individual samples used to compare with the reference template (i.e. the feature vector that resulted from the feature extraction process). From a biometric perspective, the enrolment process in feature analysis requires the existence of an enrolment sample, which is used to compute the behavioural profile of the user. This sample should contain all possible key combinations in order to effectively recognise the user based on an expected or unexpected set of author inputs. The ability to verify the user does not only have applicability in the digital forensic dominion, but could

also be used as a biometric system modality for use in transparent authentication, even if the suspect hides himself on one or more of the platforms. To this end, this research project has investigated the unification of the most discriminating features from different messaging platforms (a user profile is unified across platforms), and has involved carrying out experimentations using several of the top-most important feature tests to evaluate the viability of incorporating user profiling. It has included identifying as much genuine information from users as possible and incorporating this into a classifier; this approach has then suggested what factors can possibly be picked up. The results demonstrate a significant improvement and they show that is pragmatically better than using an individual approach for prioritised users (profiles), allowing the forensic investigator to conduct further research, and possibly reach even better performance.

However, there are some limitations, as the portability of the method needs to be further researched, and there is some limitation in the datasets. In addition, although the research has involved collecting a certain volume of data from users across main modern four platforms, further research that includes more than four platforms, and uses different methods, procedures, or sample sizes, would allow for a more extensive evaluation.

To conclude, understanding linguistic behaviour by profiling users across different messaging platforms for many different languages and trends will be crucial in the near future, as more messaging applications and systems emerge, and the linguistic feature analysis across them matures and shows greater potential for deployment across messaging platforms. It is envisaged that the ever-growing breadth of messaging platforms available to users, including terrorist organisations, could become a primary motivation for investigators and language analysts focusing on author linguistic profile trends as a means to chase them and reveal

their identities before more victims are targeted on the new modern messaging applications.

## References


[1] Abbasi, A., & Chen, H. (2005).Applying authorship analysis to extremist-group web forum messages. *Intelligent Systems, IEEE*, *20*(5), 67-75.

[2] Abbasi, A., & Chen, H. (2008).Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)*, *26*(2), 7.

[3] Afroz, S., Islam, A. C., Stolerman, A., Greenstadt, R., & McCoy, D. (2014, May). Doppelgänger finder: Taking stylometry to the underground. In *Security and Privacy (SP), 2014 IEEE Symposium on* (pp. 212-226). IEEE.

[4] Akhtar, Z., Hadid, A., Nixon, M., Tistarelli, M., Dugelay, J. L., & Marcel, S. (2017). Biometrics: In search of identity and security (Q & A). IEEE MultiMedia.

[5] Akkarapatty, N., Muralidharan, A., Raj, N. S., & Vinod, P. (2017). Dimensionality reduction techniques for text mining. In Collaborative filtering using data mining and analysis (pp. 49-72). IGI Global.

[6] Aljumily, R. (2017). Identifying Anonymous Online Message Senders: A Proposal Toward a Linguistic Fingerprint Biometric Database (LFBD). Available at SSRN 3093279.

[7] Ali, A., Shamsuddin, S. M., & Ralescu, A. L. (2015). Classification with class imbalance problem: a review. Int. J. Advance Soft Compu. Appl, 7(3), 176-204.

[8] Ali, B., & Awad, A. (2018). Cyber and physical security vulnerability assessment for IoT-based smart homes. Sensors, 18(3), 817.

[9] Ali, N., Hindi, M., & Yampolskiy, R. V. (2011, October). Evaluation of authorship attribution software on a Chat bot corpus. In *Information, Communication and Automation Technologies (ICAT), 2011 XXIII International Symposium on* (pp. 1-6). IEEE.

[10] Allison, B., & Guthrie, L. (2008, May). Authorship Attribution of E-Mail: Comparing Classifiers over a New Corpus for Evaluation. In *LREC*.

[11] Almishari, M., Oguz, E., &Tsudik, G. (2014 b, October).Fighting authorship linkability with crowdsourcing. In *Proceedings of the second edition of the ACM conference on Online social networks* (pp. 69-82). ACM.

[12] Almishari, M., Oguz, E., & Tsudik, G. (2015). Trilateral Large-Scale OSN Account Linkability Study. *arXiv preprint arXiv:1510.00783*.

[13] Almishari, M., Kaafar, D., Oguz, E., &Tsudik, G. (2014 a, November).StylometricLinkability of Tweets.In *Proceedings of the 13th Workshop on Privacy in the Electronic Society* (pp. 205-208).ACM.

[14] Altamimi, A., Clarke, N., Furnell, S., & Li, F. (2019, November). Multi-Platform Authorship Verification. In Proceedings of the Third Central European Cybersecurity Conference (pp. 1-7).

[15] Argamon, S., Šarić, M., & Stein, S. S. (2003, August). Style mining of electronic messages for multiple authorship discrimination: first results. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 475-480).ACM.

[16] Ashcroft, M., Fisher, A., Kaati, L., Omer, E., & Prucha, N. (2015, September). Detecting jihadist messages on twitter. In *2015 European Intelligence and Security Informatics Conference* (pp. 161-164). IEEE.

[17] Baayen, H., van Halteren, H., Neijt, A., & Tweedie, F. (2002, March). An experiment in authorship attribution.In *6th JADT* (pp. 29-37).

[18] Backstrom, L., & Leskovec, J. (2011, February). Supervised random walks: predicting and recommending links in social networks. In *Proceedings of the fourth ACM international conference on Web search and data mining* (pp. 635-644). ACM.

[19] Ball, James, and Paul Lewis. "Twitter and the Riots: How the News Spread." The Guardian, Guardian News and Media, 7 Dec. 2011, www.theguardian.com/uk/2011/dec/07/twitter-riots-how-news-spread (accessed 23/11/19).

[20] Barrón-Cedeno, A., Jaradat, I., Da San Martino, G., & Nakov, P. (2019). Proppy: Organizing the news based on their propagandistic content. Information Processing & Management, 56(5), 1849-1864.

[21] Belvisi, N. M. S., Muhammad, N., & Alonso-Fernandez, F. (2020, April). Forensic Authorship Analysis of Microblogging Texts Using N-Grams and Stylometric Features. In 2020 8th International Workshop on Biometrics and Forensics (IWBF) (pp. 1-6). IEEE.

[22] Bodine-Baron, E., Helmus, T. C., Magnuson, M., & Winkelman, Z. (2016). Examining ISIS support and opposition networks on Twitter. RAND Corporation Santa Monica United States.

[23] Bolle, R., Connell, J., Pankanti ,S., Ratha, N. & Senior, A. (2013). Guide to Biometrics.Springer New York.

[24] Bouchrika, I., Goffredo, M., Carter, J., & Nixon, M. (2011). On using gait in forensic biometrics. Journal of forensic sciences, 56(4), 882-889.

[25] Brocardo, M. L., Traore, I., Saad, S., &Woungang, I. (2013, May). Authorship verification for short messages using stylometry.In *Computer, Information and*

*Telecommunication Systems (CITS), 2013 International Conference on* (pp. 1-6).IEEE.

[26] Brocardo , M., Traore, I., Saad, S., &Woungang, I. (2014 a). Verifying Online User Identity using Stylometric Analysis for Short Messages. *Journal of Networks*, *9*(12), 3347-3355.

[27] Brocardo , M. L., &Traore, I. (2014 b ).Continuous authentication using micro-messages.In *Privacy, Security and Trust (PST), 2014 Twelfth Annual International Conference on* (pp.179-188). IEEE.

[28] Brocardo, M. L., Traore, I., Woungang, I., & Obaidat, M. S. (2017). Authorship verification using deep belief network systems. International Journal of Communication Systems, 30(12), e3259.

[29] *research (pp. 1-5*).Digital Government Society of North America.

[30] Chen, I., Johansson, F. D., & Sontag, D. (2018). Why is my classifier discriminatory?. In *Advances in Neural Information Processing Systems* (pp. 3539-3550).

[31] Cheng, N., Chen, X., Chandramouli, R., & Subbalakshmi, K. P. (2009, March). Gender identification from e-mails. In *Computational Intelligence and Data Mining, 2009. CIDM'09. IEEE Symposium on* (pp. 154-158). IEEE.

[32] Chen, X., Hao, P., Chandramouli, R., &Subbalakshmi, K. P. (2011).Authorship similarity detection from Email messages.In *Machine Learning and Data Mining in Pattern Recognition* .Springer Berlin Heidelberg.

[33] Chin, D. N., & Wright, W. R. (2014, July). Social Media Sources for Personality Profiling. In *UMAP Workshops*.

[34] Albon, Chris. “Feature Selection Using Random Forest.” Chris Albon, 20 Dec. 2017, chrisalbon.com/machine_learning/trees_and_forests/feature_selection_using_random_forest/ (accessed 25/11/19).

[35] Cai, Ying, Shunan Zhang, Hongke Xia, Yanfang Fan, and Haochen Zhang. "A Privacy-preserving Scheme for Interactive Messaging over Online Social Networks." IEEE Internet of Things Journal (2020).

[36] Chen, C. (2018). Infrastructure-based anonymous communication protocols in future internet architectures (Doctoral dissertation, Carnegie Mellon University).

[37] Chung, C. T., Lin, C. J., Lin, C. H., & Cheng, P. J. (2014, August). Person identification between different online social networks. In Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI)

and Intelligent Agent Technologies (IAT)-Volume 01 (pp. 94-101). IEEE Computer Society.

[38] Clarke N, Furnell M and Reynolds P (2002) Biometric Authentication for Mobile Devices. *Electronic Engineering*. In Proceeding of the 3rd Ausralian Information Warare and Security Conference, pp. 61-69.

[39] Clarke, N. L., & Furnell, S. M. (2007). Advanced user authentication for mobile devices. computers & security, 26(2), 109-119.

[40] Clarke, N., Li, F., & Furnell, S. (2017). A novel privacy preserving user identification approach for network traffic. computers & security, 70, 335-350.

[41] Clarke N (2011) *Transparent User Authentication: Biometrics, RFID and Behavioural Profiling*. Springer London.

[42] Corney, M., De Vel, O., Anderson, A., & Mohay, G. (2002). Gender-preferential text mining of e-mail discourse. In *Computer Security Applications Conference, 2002. Proceedings. 18th Annual* (pp. 282-289). IEEE.

[43] Cramer-Petersen, C. L., Christensen, B. T., & Ahmed-Kristensen, S. (2019). Empirically analysing design reasoning patterns: Abductive-deductive reasoning patterns dominate design idea generation. Design Studies, 60, 39-70.

[44] Creswell, J. W., & Poth, C. N. (2017). *Qualitative inquiry and research design: Choosing among five approaches*. Sage publications.

[45] Dargan, S., & Kumar, M. (2020). A comprehensive survey on the biometric recognition systems based on physiological and behavioral modalities. Expert Systems with Applications, 143, 113114.

[46] DeLeeuw, E. D. (2018, August). Mixed-Mode: Past, present, and future. In *Survey Research Methods* (Vol. 12, No. 2, pp. 75-89).

[47] De Vel, O., Anderson, A. M., Corney, M. W., & Mohay, G. M. (2001). Multi-topic e-mail authorship attribution forensics.

[48] Dewan, P., Kashyap, A., & Kumaraguru, P. (2014, September). Analyzing social and stylometric features to identify spear phishing emails. In *2014 APWG Symposium on Electronic Crime Research (eCrime)* (pp. 1-13). IEEE.

[49] Ding, S. H., Fung, B. C., Iqbal, F., & Cheung, W. K. (2017). Learning stylometric representations for authorship analysis. *IEEE transactions on cybernetics*, (99), 1-15.

[50] Delany, S. J., Buckley, M., & Greene, D. (2012). SMS spam filtering: methods and data. *Expert Systems with Applications*, *39*(10), 9899-9908.

[51] Elkahky, A. M., Song, Y., & He, X. (2015, May). A multi-view deep learning approach for cross domain user modeling in recommendation systems. In Proceedings of the 24th International Conference on World Wide Web (pp. 278-288). International World Wide Web Conferences Steering Committee.

[52] Farahbakhsh, R., Cuevas, A., & Crespi, N. (2016). Characterization of cross-posting activity for professional users across Facebook, Twitter and Google+. Social Network Analysis and Mining, 6(1), 33.

[53] Faraway, J. J., & Augustin, N. H. (2018). When small data beats big data. Statistics & Probability Letters, 136, 142-145.

[54] Forstall, C. W., Cavalcante, T., Theophilo, A., Shen, B., Carvalho, A. R., & Stamatatos, E. (2016). Authorship Attribution for Social Media Forensics.

[55] Fourkioti, O., Symeonidis, S., & Arampatzis, A. (2019). Language models and fusion for authorship attribution. Information Processing & Management, 56(6), 102061.

[56] Fridman, A., Stolerman, A., Acharya, S., Brennan, P., Juola, P., Greenstadt, R., & Kam, M. (2013). Decision fusion for multimodal active authentication. *IT Professional*, *15*(4), 29-33.

[57] Furnell S,. Clarke, N. (2005) Biometrics: no silver bullets. *Computer Fraud & Security* 2005(8): 9–14.

[58] Furnell, S., Rodwell, P. and Reynolds, P. (2001). "A Conceptual Security Framework to Support Continuous Subscriber Authentication in Third Generation Networks", Proceedings of Euromedia 2001.

[59] Gehrke, G., MARTELL, C., SCHEIN, A., & Anand, P. (2009). Projecting away the class imbalance problem in author attribution. *International Journal of Semantic Computing*, *3*(03), 365-382.

[60] Goga, O., Loiseau, P., Sommer, R., Teixeira, R., & Gummadi, K. P. (2015, August). On the reliability of profile matching across large online social networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1799-1808). ACM.

[61] Gorodnichy, D., Yanushkevich, S., & Shmerko, V. (2014, December). Automated border control: Problem formalization. In *Computational Intelligence in Biometrics and Identity Management (CIBIM), 2014 IEEE Symposium on* (pp. 118-125). IEEE.

[62] Green, R and Sheppard, J. "Comparing Frequency- and Style-Based Features for Twitter Author Identification," *Proc. Twenty-Sixth International Florida Artificial Intelligence Research Society Conference*, 2013.

[63] Gulati, PM, (2009). Research Management: Fundamental and Applied Research. Global India Publications, P.42

[64] Harris, L., & Harrigan, P. (2015). Social media in politics: the ultimate voter engagement tool or simply an echo chamber?. Journal of Political Marketing, 14(3), 251-283.

[65] Hanley, J., Al Mhamied, A., Cleveland, J., Hajjar, O., Hassan, G., Ives, N., ... & Hynie, M. (2018). The social networks, social support and social capital of Syrian refugees privately sponsored to settle in Montreal: Indications for employment and housing during their early experiences of integration. Canadian Ethnic Studies, 50(2), 123-148.

[66] Howedi, F., & Mohd, M. (2014). Text Classification for Authorship Attribution Using Naive Bayes Classifier with Limited Training Data. *Computer Engineering and Intelligent Systems*, *5*(4), 48-56.

[67] Hussain, A., Vatrapu, R., Hardt, D., & Jaffari, Z. A. (2014). Social data analytics tool: A demonstrative case study of methodology and software. In *Analyzing Social Media Data and Web Networks* (pp. 99-118). Palgrave Macmillan, London.

[68] Lichtenwalter, R. N., Lussier, J. T., & Chawla, N. V. (2010, July). New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 243-252). ACM.

[69] Introna, L., & Nissenbaum, H. (2010). Facial recognition technology a survey of policy and implementation issues.

[70] Iqbal, F., Khan, L. A., Fung, B., &Debbabi, M. (2010a). E-mail authorship verification for forensic investigation. In *Proceedings of the 2010 ACM Symposium on Applied Computing* (pp. 1591-1598).ACM.

[71] Iqbal, F., Binsalleeh, H., Fung, B. C., &Debbabi, M. (2010b).Mining writeprints from anonymous e-mails for forensic investigation. *Digital Investigation*, *7*(1), 56-64.

[72] Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical pattern recognition: A review. IEEE Transactions on pattern analysis and machine intelligence, 22(1), 4-37.

[73] Jain, A. K., Feng, J., & Nandakumar, K. (2010). Fingerprint matching. Computer, 43(2), 36-44.

[74] Jain, A. K., Ross, A., & Prabhakar, S. (2004). An introduction to biometric recognition. IEEE Transactions on circuits and systems for video technology, 14(1).

[75] Jain AK, Flynn P and Ross AA (2007) *Handbook of Biometrics*. Springer Science & Business Media. Available at: https://books.google.com/books?hl=en&lr=&id=WfCowMOvpioC&pgis=1 (Accessed 25/14/19).

[76] Jain, P., Kumaraguru, P., & Joshi, A. (2015, August). Other times, other values: leveraging attribute history to link user profiles across online social networks. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media* (pp. 247-255). ACM.

[77] Jawhar, J. (2016). Terrorists' Use Of The Internet: The Case Of Daesh. *Kuala Lumpur, Malaysia: The Southeast Asia Regional Centre for Counter-Terrorism (SEARCCT), Ministry of Foreign Affairs*.

[78] Jonathan, W. (2010). Essentials of Business Research: A guide to doing your research project. *London: SAGE. Kartha, CP (2006). Learning business statistics vs traditional. Business Review*, *5*, 27-33.

[79] Joo, T. M., & Teng, C. E. (2017). Impacts of social media (facebook) on human communication and relationships: A view on behavioral change and social unity. International Journal of Knowledge Content Development & Technology, 7(4), 27-50.

[80] Juola, Patrick (2006). "Authorship Attribution". Foundations and Trends in Information Retrieval Keselj, V., Peng, F., Cercone, N., & Thomas, C. (2003, August). N-gram-based author profiles for authorship attribution. In *Proceedings of the conference pacific association for computational linguistics, PACLING* (Vol. 3, pp. 255-264).

[81] Kebede, A. M., Tefrie, K. G., &Sohn, K. A. (2015, August).Anonymous Author Similarity Identification. In *IT Convergence and Security (ICITCS), 2015 5th International Conference on* (pp. 1-5). IEEE.

[82] Khanum, M., Mahboob, T., Imtiaz, W., Ghafoor, H. A., & Sehar, R. (2015). A Survey on Unsupervised Machine Learning Algorithms for Automation, Classification and Maintenance. *International Journal of Computer Applications*, *119*(13).

[83] Klaussner, C., Nerbonne, J., & Çöltekin, Ç. (2015). Finding characteristic features in stylometric analyz is. *Digital Scholarship Humanities*, fqv048. http://dx.doi.org/10.1093/llc/fqv048

[84] Koppel, M., &Schler, J. (2004, July).Authorship verification as a one-class classification problem.In *Proceedings of the twenty-first international conference on Machine learning* (p. 62).ACM.

[85] Koppel, M., Schler, J., & Argamon, S. (2013). Authorship Attribution: What's Easy and What's Hard?.

[86] Koppel, M., Argamon, S., &Shimoni, A. R. (2002).Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, *17*(4), 401-412.

[87] Koppel, M., Schler, J., & Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, *60*(1), 9-26.

[88] Koppel, M., & Schler, J. (2003, August). Exploiting stylistic idiosyncrasies for authorship attribution. In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis* (Vol. 69, p. 72).

[89] Koppel, M., Schler, J., & Bonchek-Dokow, E. (2007). Measuring differentiability: Unmasking pseudonymous Authors, Journal of Machine Learning Research, 8, 1261-1276.

[90] Korayem, M., & Crandall, D. (2013, June). De-anonymizing users across heterogeneous social computing platforms. In Seventh International AAAI Conference on Weblogs and Social Media.

[91] Kour, J., Hanmandlu, M., & Ansari, A. Q. (2011, November). Online signature verification using GA-SVM. In *Image Information Processing (ICIIP), 2011 International Conference on* (pp. 1-4). IEEE.

[92] Layton, R., Watters, P., &Dazeley, R. (2010, July).Authorship attribution for twitter in 140 characters or less. In *Cybercrime and Trustworthy Computing Workshop (CTC), 2010 Second* (pp. 1-8). IEEE.

[93] Li, J. S., Monaco, J. V., Chen, L. C., & Tappert, C. C. (2014, November). Authorship Authentication Using Short Messages from Social Networking Sites. In e-Business Engineering (ICEBE), 2014 IEEE 11th International Conference on (pp. 314-319). IEEE.

[94] Li, J. S. (2015). An investigation of authorship authentication in short messages from a social networking site.

[95] Liu, S., Wang, S., Zhu, F., Zhang, J., & Krishnan, R. (2014, June). Hydra: Large-scale social identity linkage via heterogeneous behavior modeling. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data* (pp. 51-62). ACM.

[96] Mac, R. (2020) Tesla's billionaire CEO fought a lawsuit over his own bad behavior and won Available at https://www.buzzfeednews.com/article/ryanmac/elon-musk-cant-lose

[97] MacLeod, N., & Grant, T. (2012).Whose Tweet?Authorship analysis of micro-blogs and other short-form messages.210-224.

[98] Madigan, D., Genkin, A., Lewis, D. D., Argamon, S., Fradkin, D., & Ye, L. (2005, June). Author identification on the large scale. In Proc. of the Meeting of the Classification Society of North America (Vol. 13).

[99] Maheswaran, M., Ali, B., Ozguven, H., & Lord, J. (2010). Online identities and social networking. In *Handbook of Social Network Technologies and Applications* (pp. 241-267). Springer US.

[100] Maitra, P., Ghosh, S., & Das, D. (2016). Authorship Verification-An Approach based on Random Forest. *arXiv preprint arXiv:1607.08885*.

[101] Mariappan, P., Padhmavathi, B., & Teja, T. S. (2016). Digital Forensic and Machine Learning. In *Combating Security Breaches and Criminal Activity in the Digital Sphere* (pp. 141-156). IGI Global.

[102] Marques, O. (2016). Social Networks. In *Innovative Technologies in Everyday Life* (pp. 31-44). Springer International Publishing.

[103] McBride, K. A., MacMillan, F., George, E. S., & Steiner, G. Z. (2019). The Use of Mixed Methods in Research. *Handbook of Research Methods in Health Social Sciences*, 695-713.

[104] Meghanathan, N. (2018). Biometric Systems for User Authentication. In Computer and Network Security Essentials (pp. 317-335). Springer, Cham.

[105] Mikros, G. K. (2012). Authorship attribution and gender identification in Greek blogs. *Methods and Applications of Quantitative Linguistics*, *21*, 21-32.

[106] Mikros, G. K., & Argiri, E. K. (2007, July). Investigating Topic Influence in Authorship Attribution. In PAN.

[107] Monaco, J. V., Stewart, J. C., Cha, S. H., & Tappert, C. C. (2013, September). Behavioral biometric verification of student identity in online course assessment and authentication of authors in literary works. In *Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on* (pp. 1-8). IEEE.

[108] Mukherjee, A., Venkataraman, V., Liu, B., & Glance, N. (2013, June). What yelp fake review filter might be doing?. In *Seventh international AAAI conference on weblogs and social media*.

[109] Nagaprasad, S., Reddy, T. R., Reddy, P. V., Babu, A. V., & VishnuVardhan, B. (2015). Empirical evaluations using character and word n-grams on authorship attribution for Telugu text. In Intelligent Computing and Applications (pp. 613-623). Springer, New Delhi.

[110] Narayanan, A., Paskov, H., Gong, N. Z., Bethencourt, J., Stefanov, E., Shin, E. C. R., & Song, D. (2012). On the feasibility of internet-scale author identification. In *Security and Privacy (SP), 2012 IEEE Symposium on* (pp. 300-314). IEEE.

[111] Neal, T., Sundararajan, K., Fatima, A., Yan, Y., Xiang, Y., & Woodard, D. (2017). Surveying stylometry techniques and applications. ACM Computing Surveys (CSUR), 50(6), 1-36.

[112] Nelufule, N. N. (2014). *Combining Multiple Iris Matchers using Advanced Fusion Techniques to Enhance Iris Matching Performance*.

[113] Newton, Casey. “Facebook's Role in the French Protests Has Polarized Observers.” The Verge, The Verge, 11 Dec. 2018, www.theverge.com/2018/12/11/18135273/yellow-vest-facebook-france-protests (Accessed: 23/11/2019).

[114] Nguyen, M. H., & De la Torre, F. (2010). Optimal feature selection for support vector machines. *Pattern recognition*, *43*(3), 584-591.

[115] Nirkhi, S. (2019). Evaluation of Classifiers for Detection of Authorship Attribution. In Computational Intelligence: Theories, Applications and Future Directions-Volume I (pp. 227-236). Springer, Singapore.

[116] Nirkhi, S. M., Dharaskar, R. V., &Thakre, V. M. (2012, June).Analysis of online messages for identity tracing in cybercrime investigation. In *Cyber Security, Cyber Warfare and Digital Forensic (CyberSec), 2012 International Conference on* (pp. 300-305). IEEE.

[117] Nirkhi, S., &Dharaskar, R. V. (2013).Comparative study of authorship identification techniques for cyber forensics analysis. *arXiv preprint arXiv:1401.6118*.

[118] Nirkhi, S. M., Dharaskar, R. V., &Thakare, V. M. (2015).Authorship Identification using Generalized Features and Analysis of Computational Method. *Transactions on Machine Learning and Artificial Intelligence*, *3*(2), 41.

[119] Novak, J., Raghavan, P., & Tomkins, A. (2004).Anti-aliasing on the web.In*Proceedings of the 13th international conference on World Wide Web*. New York, NY, USA: ACM, 2004, pp. 30–39.

[120] Obaidat, M. S., Traore, I., & Woungang, I. (Eds.). (2019). Biometric-Based Physical and Cybersecurity Systems (Vol. 368). Springer.

[121] Olsson J. 2004. Forensic linguistics: an introduction to language, crime and the law.London/New York, Continuum, Pp. xiii + 269.

[122] Orebaugh, A. (2006, October). An Instant Messaging Intrusion Detection System Framework: Using character frequency analysis for authorship identification and validation. In *Carnahan Conferences Security Technology, Proceedings 2006 40th Annual IEEE International* (pp. 160-172).IEEE.

[123] Ottoni, R., Las Casas, D., Pesce, J. P., Meira Jr, W., Wilson, C., Mislove, A., & Almeida, V. (2014, May). Of pins and tweets: Investigating how users behave across image-and text-based social networks. In Eighth international aaai conference on weblogs and social media.

[124] Overdorf, R., Dutko, T., & Greenstadt, R. (2014). Blogs and Twitter Feeds: A Stylometric Environmental Impact Study.

[125] Page, R., Barton, D., Unger, J. W., & Zappavigna, M. (2014). *Researching language and social media: A student guide*. Routledge.

[126] Pavelec, D., Oliveira, L. S., Justino, E., Neto, F. N., & Batista, L. V. (2009, June). Compression and stylometry for author identification. In *Neural Networks, 2009. IJCNN 2009. International Joint Conference on* (pp. 2445-2450). IEEE.

[127] Ragel, R., Herath, P., &Senanayake, U. (2013, December). Authorship detection of SMS messages using unigrams. In *Industrial and Information Systems (ICIIS), 2013 8th IEEE International Conference on* (pp. 387-392). IEEE.

[128] Rainie, H., Anderson, J. Q., & Albright, J. (2017). The future of free speech, trolls, anonymity and fake news online.

[129] Rappoport, R. S. O. T. A., & Koppel, M. (2013). Authorship attribution of micro-messages.

[130] Reichart Smith, L., Smith, K. D., & Blazka, M. (2017). Follow Me, What's the Harm: Considerations of Catfishing and Utilizing Fake Online Personas on Social Media. J. Legal Aspects Sport, 27, 32

[131] Reuters Institute Digital News Report (2019). [ONLINE] Available at: from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2019-06/DNR_2019_FINAL_1.pdf. (Accessed: 25/11/2019).

[132] Robinson,D. (2016). Text analysis of Trump's tweets confirms he writes only the (angrier) Android half. Available online at http://varianceexplained.org/r/trump-tweets/ Last access 27/03/2017.

[133] Rocha, A., Scheirer, W. J., Forstall, C. W., Cavalcante, T., Theophilo, A., Shen, B., ... & Stamatatos, E. (2016). Authorship attribution for social media forensics. IEEE Transactions on Information Forensics and Security, 12(1), 5-33.

[134] Ross A (2007) An Introduction to Multibiometrics. *the 15th European Signal Processing Conference (EUSIPCO)*. Poznan, Poland, 20–24.

[135] Russell, M. A. (2013). Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More. " O'Reilly Media, Inc.". The guardian (2011) available online at https://www.theguardian.com/world/2016/jan/25/egypt-5-years-on-was-it-ever-a-social-media-revolution (Accessed: 25/11/2019).

[136] Saevanee, H., Clarke, N., & Furnell, S. (2011, September). SMS linguistic profiling authentication on mobile device. In 2011 5th International Conference on Network and System Security (pp. 224-228). IEEE.

[137] Saevanee, H., Clarke, N., Furnell, S., & Biscione, V. (2015). Continuous user authentication using multi-modal biometrics. *Computers & Security*, *53*, 234-246.

[138] Sajjad, M., Nasir, M., Muhammad, K., Khan, S., Jan, Z., Sangaiah, A. K., ... & Baik, S. W. (2020). Raspberry Pi assisted face recognition framework for enhanced law-enforcement services in smart cities. Future Generation Computer Systems, 108, 995-1007.

[139] Sanderson, C., & Guenter, S. (2006, July). Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (pp. 482-491). Association for Computational Linguistics.

[140] Salim, B. (2012). An application of UTAUT model for acceptance of social media in Egypt: A statistical study. International Journal of Information Science, 2(6), 92-105.

[141] Sankaran, A., Jain, A., Vashisth, T., Vatsa, M., & Singh, R. (2017). Adaptive latent fingerprint segmentation using feature selection and random decision forest classification. Information Fusion, 34, 1-15.

[142] Schonlau, M., Guenther, N., & Sucholutsky, I. (2017). Text mining with n-gram variables. The Stata Journal, 17(4), 866-881.

[143] Schultz, Jeff. “Micro Focus Blog.” How Much Data Is Created on the Internet Each Day? | Micro Focus Blog, 8 June 2019, blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day/. (Accessed: 25/11/2019).

[144] Sharma, E. K., & Banga, V. K(2013). Biometric Security Issues: A Review.

[145] Shirish, T. S. (2013). *Research methodology in education*. Lulu. com.

[146] Siham, O. & Halim, S. (2012). Authorship Attribution of Ancient Texts Written by Ten Arabic Travelers Using a SMO-SVM Classifier, *The 2nd International Conference on Communications and Information Technology (ICCIT): Digital Information Management, Hammamey,* 44-47.

[147] Silva, R. S., Laboreiro, G., Sarmento, L., Grant, T., Oliveira, E., & Maia, B. (2011). ‘twazn me!!!;(’automatic authorship analysis of micro-blogging messages. In *Natural Language Processing and Information Systems* (pp. 161-168). Springer Berlin Heidelberg.

[148] Silverman, B. W. (2018). *Density estimation for statistics and data analysis*. Routledge

[149] Silverman, D. (2016). Introducing qualitative research. Qualitative research, 3-14.

[150] Singh, S. (2018). Forensic and Automatic Speaker Recognition System. International Journal of Electrical and Computer Engineering, 8(5), 2804.

[151] Singh, P. K., Vivek, K. S., & Kodimala, S. (2017, October). Stylometric analysis of E-mail content for author identification. In Proceedings of the 1st International Conference on Internet of Things and Machine Learning (p. 61). ACM.

[152] Smith, K. (2019). 126 Amazing Social Media Statistics and Facts. [ONLINE] Available at: https://www.brandwatch.com/blog/amazing-social-media-statistics-and-facts/ (Accessed: 25/11/2019).

[153] Stamatatos, E. (2008). Author identification: Using text sampling to handle the class imbalance problem. *Information Processing & Management*, *44*(2), 790-799.

[154] Stamatatos, E., Fakotakis, N., &Kokkinakis, G. (2001). Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, *35*(2), 193-214.

[155] Stamatatos, E. (2007, September). Author identification using imbalanced and limited training texts. In *18th International Workshop on Database and Expert Systems Applications (DEXA 2007)* (pp. 237-241). IEEE.

[156] Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, *60*(3), 538-556.

[157] Stańczyk, U. (2016). The class imbalance problem in construction of training datasets for authorship attribution. In *Man-Machine Interactions 4* (pp. 535-547). Springer, Cham.

[158] Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., & Demirbas, M. (2010, July). Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (pp. 841-842). ACM.

[159] Stringhini, G., Kruegel, C., & Vigna, G. (2010, December). Detecting spammers on social networks. In *Proceedings of the 26th annual computer security applications conference* (pp. 1-9). ACM.

[160] Steyvers, M., Smyth, P., Rosen-Zvi, M., & Griffiths, T. (2004, August). Probabilistic author-topic models for information discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 306-315). ACM.

[161] Tan, R. H. R., & Tsai, F. S. (2010, October). Authorship identification for online text. In *Cyberworlds (CW), 2010 International Conference on* (pp. 155-162). IEEE.

[162] Ensor, Josie. “Husband Killed and Burnt Wife, Then Took World Trip on Her Cash.” The Telegraph, Telegraph Media Group, 11 May 2013, www.telegraph.co.uk/news/uknews/crime/10050579/Husband-killed-and-burnt-wife-then-took-world-trip-on-her-cash.html (Accessed: 23/11/2019).

[163] Torgo, L. (2011). Data mining with R: learning with case studies. Chapman and Hall/CRC.

[164] Torgo, L. (2016). Data Mining with R: Learning with Case Studies. CRC Press.

[165] Tsesis, A. (2017). Terrorist speech on social media. Vand. L. Rev., 70, 651.

[166] Shearlaw, Maeve. “Egypt Five Years on: Was It Ever a 'Social Media Revolution'?” The Guardian, Guardian News and Media, 25 Jan. 2016, www.theguardian.com/world/2016/jan/25/egypt-5-years-on-was-it-ever-a-social-media-revolution (Accessed: 23/11/2019).

[167] Roberts, Rachel. “Online Trolling Laws under Consideration Following Abuse of MPs.” The Independent, Independent Digital News and Media, 24 July 2017, www.independent.co.uk/news/uk/politics/trolling-laws-online-abuse-mps-under-consideration-lord-bew-a7857891.html (accessed 23/11/2019).

[168] Peltier, Elian, and Adam Satariano. “After Yellow Vests Come Off, Activists in France Use Facebook to Protest and Plan.” The New York Times, The New York Times, 14 Dec. 2018, www.nytimes.com/2018/12/14/technology/facebook-france-yellow-vests.html (Accessed: 23/11/2019).

[169] Tonkin, E., Pfeiffer, H. D., & Tourte, G. (2012). Twitter, information sharing and the London riots?. *Bulletin of the American Society for Information Science and Technology*, *38*(2), 49-57.

[170] Vosoughi, S., Zhou, H., & Roy, D. (2015, December). Digital stylometry: Linking profiles across social networks. In *International Conference on Social Informatics* (pp. 164-177). Springer, Cham.

[171] Vorobeva, A. A. (2016, April). Examining the performance of classification algorithms for imbalanced data sets in web author identification. In *Open Innovations Association and Seminar on Information Security and Protection of Information Technology (FRUCT-ISPIT), 2016 18th Conference of* (pp. 385-390). IEEE.

[172] Ward, Antonia. “ISIS's Social Media Use Poses a Threat to Stability in the Middle East and Africa.” RAND Corporation, 11 Dec. 2018, www.rand.org/blog/2018/12/isiss-use-of-social-media-still-poses-a-threat-to-stability.html (Accessed: 23/11/2019).

[173] Wright, D. (2014). *Stylistics versus Statistics: A corpus linguistic approach to combining techniques in forensic authorship analysis using Enron Emails* (Doctoral dissertation, University of Leeds).

[174] Weir, G. R., Toolan, F., & Smeed, D. (2011). The threats of social networking: Old wine in new bottles?. *Information security technical report*, *16*(2), 38-43.

[175] Yule, G. U. (1938). On sentence length as a statistical characteristic of style in prose.*Biometrika, 30*, 363-390.

[176] Yadron, Danny. “Twitter Deletes 125,000 Isis Accounts and Expands Anti-Terror Teams.” The Guardian, Guardian News and Media, 5 Feb. 2016,

www.theguardian.com/technology/2016/feb/05/twitter-deletes-isis-accounts-terrorism-online. (Accessed: 25/11/2019).

[177] Yule, G. U. (1944). The statistical study of literary vocabulary. Cambridge, UK: Cambridge University Press.

[178] Zafarani, R., & Liu, H. (2009). Connecting Corresponding Identities across Communities. *ICWSM*, *9*, 354-357.

[179] Zafarani, R., & Liu, H. (2013, August).Connecting users across social media sites: A behavioral-modeling approach.In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 41-49).ACM.

[180] Zheng, R., Li, J., Chen, H., & Huang, Z. (2006).A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, *57*(3), 378-393.

# Appendix A – Attendance of International Corpus linguistics Conference

One of the most important international conferences specialized in Cybersecurity of Corpus Linguistics Corpus.

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

**CL 2019**
International Corpus Linguistics Conference

Cardiff, Wales, UK
July 22-26, 2019

## Certificate of Attendance

This is to certify that

# Abdulaziz Altamimi

attended the International Corpus Linguistics Conference (CL2019)
at Cardiff University, July 2019

**Dr Dawn Knight**
Chair of the CL2019
Organising Committee

**Date**: 26/7/19

# Appendix B - List of Stylometric Features Used

| Features | List of features |
|---|---|
| Character-based features<br>(F1-F50) | Feature 1: number of characters.<br>Feature 2: number of alphabets.<br>Feature 3: number of uppercase characters<br>Feature 4-29: number of alphabet a-z. Feature 30-50: number of special character "~ @ # $ % ^ & * - _ = + > < [] { } / \\ \|". |
| Syntactic Features<br>(F51-F208) | Feature 51-58: number of punctuation ", . ? ! : ; \" ' " .<br>Feature 59-208: Function words:<br>"a, about, above, after, all, although, am, among, an, and, another,any,anybody, anyone, anything, are, around, as, at, be, because, before, behind, below, beside, between, both, but, by, can, cos, do, down, each, either, enough, every, everybody, everyone, everything, few,following, for, from, have, he, her, him, I, if, in, including, inside, into, is, it, its, latter, less, like, little, lots, many, me, more, most, much, my, need, neither, no, nobody, none, nor, nothing, of, off, on, once, one, onto, opposite, or, our, outside, over, own, past, per, plenty, plus, regarding, same,several, she, should, since, so, some, somebody, someone, something, such, than, that, he, their, them, these, they, this, those, though, through, till, to, toward, towards, under, unless, unlike, until, up, upon, us, used, via, we, what, whatever, when, where, whether, which, while, who, whoever, whom, whose, will, with, within, without, worth, would, yes, you, your". |
| Structural Features (209) | Feature 209: Total number of sentences |
| Word-based features<br>(F210- F227) | Feature 210: Total number of words.<br>Feature 211: Total number of short words (less than four characters).<br>Feature 212: Average word length.<br>Feature 213: Average sentence length in terms of character. Feature 214: Average sentence length in terms of word. Feature 215: Number of words with 1 char. Feature 216: Number of words with 2 chars.<br>Feature 217: Number of words with 3 chars.<br>Feature 218: Number of words with 4 chars.<br>Feature 219:Number of words with 5 chars.<br>Feature 220: Number of words with 6 chars.<br>Feature 221: Number of words with 7 Chars.<br>Feature 222: Number of words with 8 chars.<br>Feature 223: Number of words with 9 chars.<br>Feature 224:Number of words with 10 chars.<br>Feature 225: Number of words with 11 chars.<br>Features 226: Number of words with 12 chars. |

| | |
|---|---|
| | Features 227: Number of words with more than 12 chars. |
| Social Network Specific features (F228-F233) | Feature 228: Frequency of a happy face ":)". Feature 229: Frequency of a sad face ":(". Feature 230: Frequency of "LOL". Feature 231: Frequency of missing an uppercase letter when starting sentence. Feature 232: Frequency of missing a period or other punctuation to sentence. Feature 233: Frequency of missing the word "I" or "We" in a sentence. |
| Emotional specific features (F233-F275) | 234.😁 grinning face. 235.😀 Grinning Face With Smiling Eyes. 236.😂 Face With Tears of Joy. 237.😃 Smiling Face With Open Mouth. 238.👍 Thumbs Up. 239.😉 Winking face. 240.Hugging face. 241.☺Smiling face. 242.😄 Smiling Face With Open Mouth & Closed Eyes. 243.😊 Smiling Face With Open Mouth & Smiling Eyes. 244.Face With Stuck-Out Tongue. 245.😝 Face With Stuck-Out Tongue & Winking Eye. 246.😜 Face With Stuck-Out Tongue & Closed Eyes. 247.😒 Persevering Face. 248.😓 Disappointed but Relieved Face. 249.☹.Frowning Face. 250.😭 Loudly Crying Face. 251.Slightly Frowning Face. 252.Anguished Face. 253.😕 Weary Face. 254.😵 Dizzy Face. 255.😠 Angry Face. 256.😢 Crying Face. 257.💓 beating heart. 258.💋 kiss. 259.🙏 pray. 260.raised hands with fingers. 261.✌. 262.💔 broken heart. 263.💑 couple with heart. 264.👎 thumbs down. 265.😍 heart eyes. 266.rofl. 267.Face With Rolling Eyes. 268. Slightly Smiling Face. 269.😔 Tired Face. 270.😪 Sleepy face. 271.😈 Smiling Face With Horns. 272.Grimacing Face. 273.😕 Confounded Face. 274.❤ heart. 275.👌 ok hand. |

# Appendix C – Ethical Approval, Consent Form and Information Sheet (Data Collection)

RESEARCH WITH PLYMOUTH UNIVERSITY

12 October 2017

CONFIDENTIAL
Abdulaziz Altamimi
School of Computing, Electronics and Mathematics

Dear Abdulaziz

***Ethical Approval Application***

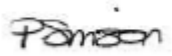
Thank you for submitting the ethical approval form and details concerning your project:

***Author Identification of Text Limited Messages***

I am pleased to inform you this has been approved subject to the following amendment:

- Section 5.3 states that: "The participants will be given the consent at the beginning of the experiment session, should they wish to carry out the study, ensuring they understand that they cannot withdraw from the experiment at any time up until the end of their participation." The words "cannot withdraw" should be "can withdraw"

Kind regards



Paula Simson
Secretary to Faculty Research Ethics Committee

Faculty of Science and Engineering
Plymouth University
Drake Circus
PL4 8AA

T +44 (0) 1752 584 584
F +44 (0) 1752 584 540
W www.plymouth.ac.uk

Mrs Jayne Brenen
Head of Faculty Operations

***SAMPLE SELF-CONSENT FORM***

**PLYMOUTH UNIVERSITY**

**FACULTY OF SCIENCE AND ENGINEERING**

**Human Ethics Committee Sample Consent Form**

CONSENT TO PARICIPATE IN RESEARCH PROJECT / PRACTICAL STUDY

---

Name of Principal Investigator

Abdulaziz Altamimi

---

Title of Research

Author Identification of Text Limited Messages.

---

Brief statement of purpose of work

1-To explore and investigate whether a single profile can be used in author identification system. It is a series of experiments to determine to what degree we can build a unified single profile across messaging systems.

2- To understand and investigate how much text required in order to achieve that.

---

The objectives of this research have been explained to me.

I understand that I have the right to withdraw at any point up to the completion of the data collection stage, at which point, due to anonymization, it will be no longer possible to identify my data from others.

I understand that my anonymity is guaranteed, unless I expressly state otherwise.

I understand that the Principal Investigator of this work will have attempted, as far as possible, to avoid any risks, and that safety and health risks will have been separately assessed by appropriate authorities (e.g. under COSHH regulations)

Under these circumstances, I agree to participate in the research.

Name: .....

Signature: ..... Date: .....

***SAMPLE INFORMATION SHEET FOR ADULT***

**PLYMOUTH UNIVERSITY**

**FACULTY OF SCIENCE AND ENGINEERING**

RESEARCH INFORMATION SHEET

---

**Name of Principal Investigator**

Abdulaziz Altamimi

---

**Title of Research**

Author Identification of Text Limited Messages.

**Aim of research**

Author identification is a technique that can be used to verify individuals based upon the composition of messages. Currently, however, such approaches need to be developed on a per-system basis (i.e. the profile used to identify an individual from instant massager would be different to email). The purpose of this experiment is to:

1- Determine to what degree a unified single profile can be used across messaging systems.

2- Understand and investigate how much text required in order to achieve that.

**Description of procedure**

Participants will be asked to sit in front of a computer machine and perform a set of logins to provide access to (up to 4) messaging systems: Facebook, Twitter, e-mail and SMS. A tool will extract their data and then calculate the necessary features required. In this manner, the highly private messages are not stored or used directly and the researcher simply removes the feature sets (which do not contain any information that could be used to recreate the original message). The figure below is an illustration of the process and the resulting data that is captured.

The output file is a set of features, the file contains no private information, and it is not possible to go from these features back to original content. It is only one way process. The scrapped data will be deleted and removed from hard drive during the session which the student present, and not taken forward at all. When the user finished his session there is no way the researcher to go back to account, there is no way to access to data, no relationship and no information about message itself.

**Description of risks**

The collected features will be treated confidentially and the data will be delated and features will be anonymous during the collection and is extracted.

**Benefits of proposed research**

This research is useful because a need exists to be able to identify the ownership of messages sent across these systems. Relying on just the account details to simply identify

# Appendix D - An automated Developed Feature Extractor Code

```
import org.apache.commons.lang3.ArrayUtils;

import java.text.BreakIterator;
import java.util.*;
import java.util.regex.Matcher;
import java.util.regex.Pattern;
/**
 * 25-10-2017
 */
public class FeaturesExtractor {
    /**
     * This method generates features vector for all given input texts.
     *
     * @param texts Input texts to be analyzed.
     * @return List of feature vectors.
     */
    public List<float[]> extractList(List<String> texts){
        List<float[]> result = new ArrayList<float[]>();
        for(String text : texts){
            if (text != null){
                result.add(extract(text));
            }

        }
        return result;
    }
    /**
     * This method extracts features from a single text.
     * @param text Input text to be analyzed.
     * @return Features vector as an array of float numbers.
     */
    public float[] extract(String text){
// split text by any punctuation or space into words
String[] words = text.split ("[^\\w']+");
// count uppercase letters
        int uppercase = 0;
        int lowercase = 0;
        for(int i = 0; i < text.length(); i++){
            if(Character.isUpperCase(text.charAt(i))) uppercase++;
        }
        for(int i = 0; i < text.length(); i++){
            if(Character.isLowerCase(text.charAt(i))) lowercase++;
        }
        float[] firstTriad = new float[3];
        firstTriad[0] = text.replaceAll(" ", "").length();
        firstTriad[1] = uppercase+lowercase;;
        firstTriad[2] = uppercase;
// join results of each analysis part into a single vector
        float[]          res          =          ArrayUtils.addAll(firstTriad,
ArrayUtils.addAll(getCharsCount(text),   ArrayUtils.addAll(getWordCount(words),
ArrayUtils.addAll(getLenghtsCount(text, words), getEmojiCount(text)))));
        return res;
    }
    /**
     * This method is used to count number of occurrences of each character in
thegiven text.
     * @param text Input text to be analyzed.
     * @return Array of numbers of occurrences of each letter, punctuation and
special
     * characters.
     */
private float[] getCharsCount(String text){
        char charArray[] = text.toLowerCase().toCharArray();

        float[] count = new float[55];
```

```
for(int i = 0; i < count.length; i++){
    count[i] = 0;
}
/*
 */
for (char c : charArray) {
    switch (c) {
        case 'a':
            count[0]++;
            break;
        case 'b':
            count[1]++;
            break;
        case 'c':
            count[2]++;
            break;
        case 'd':
            count[3]++;
            break;
        case 'e':
            count[4]++;
            break;
        case 'f':
            count[5]++;
            break;
        case 'g':
            count[6]++;
            break;
        case 'h':
            count[7]++;
            break;
        case 'i':
            count[8]++;
            break;
        case 'j':
            count[9]++;
            break;
        case 'k':
            count[10]++;
            break;
        case 'l':
            count[11]++;
            break;
        case 'm':
            count[12]++;
            break;
        case 'n':
            count[13]++;
            break;
        case 'o':
            count[14]++;
            break;
        case 'p':
            count[15]++;
            break;
        case 'q':
            count[16]++;
            break;
        case 'r':
            count[17]++;
            break;
        case 's':
            count[18]++;
            break;
        case 't':
            count[19]++;
            break;
        case 'u':
            count[20]++;
```

```
            break;
        case 'v':
            count[21]++;
            break;
        case 'w':
            count[22]++;
            break;
        case 'x':
            count[23]++;
            break;
        case 'y':
            count[24]++;
            break;
        case 'z':
            count[25]++;
            break;
        case '~':
            count[26]++;
            break;
        case '@':
            count[27]++;
            break;
        case '#':
            count[28]++;
            break;
        case '$':
            count[29]++;
            break;
        case '%':
            count[30]++;
            break;
        case '^':
            count[31]++;
            break;
        case '&':
            count[32]++;
            break;
        case '*':
            count[33]++;
            break;
        case '-':
            count[34]++;
            break;
        case '_':
            count[35]++;
            break;
        case '=':
            count[36]++;
            break;
        case '+':
            count[37]++;
            break;
        case '>':
            count[38]++;
            break;
        case '<':
            count[39]++;
            break;
        case '[':
            count[40]++;
            break;
        case ']':
            count[41]++;
            break;
        case '{':
            count[42]++;
            break;
        case '}':
            count[43]++;
            break;
```

```
                case '/':
                    count[44]++;
                    break;
                    count[45]++;
                    break;
                case '|':
                    count[46]++;
                    break;
                case ',':
                    count[47]++;
                    break;
                case '.':
                    count[48]++;
                    break;
                case '?':
                    count[49]++;
                    break;
                case '!':
                   count[50]++;
                    break;
                case ':':
                    count[51]++;
                    break;
                case ';':
                    count[52]++;
                    break; case '"':
                    count[53]++;
                    break;
                case '"':
                    count[53]++;
                    break;
                case '\':
                    count[54]++;
                    break;
        }
        return count;
    }
/**
     * This method is used to count number of 'function words' words in the
given text.
     * @param words Array of words of the input text.
     * @return Array with number of occurrences of interested words.
     */
    private float[] getWordCount(String[] words){
        float[] result = new float[150];

        for(int i = 0; i < result.length; i++){
            result[i] = 0;}
        for(String s : words) {
            if (s.equals("a")) {
                result[0]++;
                continue;
            }
            if (s.equals("about")) {
                result[1]++;
                continue;
            }
            if (s.equals("above")) {
                result[2]++;
                continue;
            }
            if (s.equals("after")) {
                result[3]++;
                continue;
            }
            if (s.equals("all")) {
                result[4]++;
                continue;
            }
            if (s.equals("although")) {
```

```
        result[5]++;
        continue;
    }
    if (s.equals("am")) {
        result[6]++;
        continue;
    }
    if (s.equals("among")) {
        result[7]++;
        continue;
    }
    if (s.equals("an")) {
        result[8]++;
        continue;
    }
    if (s.equals("and")) {
        result[9]++;
        continue;
    }
    if (s.equals("another")) {
        result[10]++;
        continue;
    }
    if (s.equals("any")) {
        result[11]++;
        continue;
    }
    if (s.equals("anybody")) {
        result[12]++;
        continue;
    }
    if (s.equals("anyone")) {
        result[13]++;
        continue;
    }
    if (s.equals("anything")) {
        result[14]++;
        continue;
    }
    if (s.equals("are")) {
        result[15]++;
        continue;
    }
    if (s.equals("around")) {
        result[16]++;
        continue;
    }
    if (s.equals("as")) {
        result[17]++;
        continue;
    }
    if (s.equals("at")) {
        result[18]++;
        continue;
    }
    if (s.equals("be")) {
        result[19]++;
        continue;
    }
    if (s.equals("because")) {
        result[20]++;
        continue;
    }
    if (s.equals("before")) {
        result[21]++;
        continue;
    }
    if (s.equals("behind")) {
        result[22]++;
        continue;
```

```
}
if (s.equals("below")) {
    result[23]++;
    continue;
}
if (s.equals("beside")) {
    result[24]++;
    continue;
}
if (s.equals("between")) {
    result[25]++;
    continue;
}
if (s.equals("both")) {
    result[26]++;
    continue;
}
if (s.equals("but")) {
    result[27]++;
    continue;
}
if (s.equals("by")) {
    result[28]++;
    continue;
}
if (s.equals("can")) {
    result[29]++;
    continue;
}
if (s.equals("cos")) {
    result[30]++;
    continue;
}
if (s.equals("do")) {
    result[31]++;
    continue;
}
if (s.equals("down")) {
    result[32]++;
    continue;
}
if (s.equals("each")) {
    result[33]++;
    continue;
}
if (s.equals("either")) {
    result[34]++;
    continue;
}
if (s.equals("enough")) {
    result[35]++;
    continue;
}
if (s.equals("every")) {
    result[36]++;
    continue;
}
if (s.equals("everybody")) {
    result[37]++;
    continue;
}
if (s.equals("everyone")) {
    result[38]++;
    continue;
}
if (s.equals("everything")) {
    result[39]++;
    continue;
}
if (s.equals("few")) {
```

```
        result[40]++;
        continue;
    }
    if (s.equals("following")) {
        result[41]++;
        continue;
    }
    if (s.equals("for")) {
        result[42]++;
        continue;
    }
    if (s.equals("from")) {
        result[43]++;
        continue;
    }
    if (s.equals("have")) {
        result[44]++;
        continue;
    }
    if (s.equals("he")) {
        result[45]++;
        continue;
    }
    if (s.equals("her")) {
        result[46]++;
        continue;
    }
    if (s.equals("him")) {
        result[47]++;
        continue;
    }
    if (s.equals("I")) {
        result[48]++;
        continue;
    }
    if (s.equals("if")) {
        result[49]++;
        continue;
    }
    if (s.equals("in")) {
        result[50]++;
        continue;
    }
    if (s.equals("including")) {
        result[51]++;
        continue;
    }
    if (s.equals("inside")) {
        result[52]++;
        continue;
    }
    if (s.equals("into")) {
        result[53]++;
        continue;
    }
    if (s.equals("is")) {
        result[54]++;
        continue;
    }
    if (s.equals("it")) {
        result[55]++;
        continue;
    }
    if (s.equals("its")) {
        result[56]++;
        continue;
    }
    if (s.equals("latter")) {
        result[57]++;
        continue;
```

```
}
if (s.equals("less")) {
    result[58]++;
    continue;
}
if (s.equals("like")) {
    result[59]++;
    continue;
}
if (s.equals("little")) {
    result[60]++;
    continue;
}
if (s.equals("lots")) {
    result[61]++;
    continue;
}
if (s.equals("many")) {
    result[62]++;
    continue;
}
if (s.equals("me")) {
    result[63]++;
    continue;
}
if (s.equals("more")) {
    result[64]++;
    continue;
}
if (s.equals("most")) {
    result[65]++;
    continue;
}
if (s.equals("much")) {
    result[66]++;
    continue;
}
if (s.equals("my")) {
    result[67]++;
    continue;
}
if (s.equals("need")) {
    result[68]++;
    continue;
}
if (s.equals("neither")) {
    result[69]++;
    continue;
}
if (s.equals("no")) {
    result[70]++;
    continue;
}
if (s.equals("nobody")) {
    result[71]++;
    continue;
}
if (s.equals("none")) {
    result[72]++;
    continue;
}
if (s.equals("nor")) {
    result[73]++;
    continue;
}
if (s.equals("nothing")) {
    result[74]++;
    continue;
}
if (s.equals("of")) {
```

```
        result[75]++;
        continue;
    }
    if (s.equals("off")) {
        result[76]++;
        continue;
    }
    if (s.equals("on")) {
        result[77]++;
        continue;
    }
    if (s.equals("once")) {
        result[78]++;
        continue;
    }
    if (s.equals("one")) {
        result[79]++;
        continue;
    }
    if (s.equals("onto")) {
        result[80]++;
        continue;
    }
    if (s.equals("opposite")) {
        result[81]++;
        continue;
    }
    if (s.equals("or")) {
        result[82]++;
        continue;
    }
    if (s.equals("our")) {
        result[83]++;
        continue;
    }
    if (s.equals("outside")) {
        result[84]++;
        continue;
    }
    if (s.equals("over")) {
        result[85]++;
        continue;
    }
    if (s.equals("own")) {
        result[86]++;
        continue;
    }
    if (s.equals("past")) {
        result[87]++;
        continue;
    }
    if (s.equals("per")) {
        result[88]++;
        continue;
    }
    if (s.equals("plenty")) {
        result[89]++;
        continue;
    }
    if (s.equals("plus")) {
        result[90]++;
        continue;
    }
    if (s.equals("regarding")) {
        result[91]++;
        continue;
    }
    if (s.equals("same")) {
        result[92]++;
        continue;
```

```
}
if (s.equals("several")) {
    result[93]++;
    continue;
}
if (s.equals("she")) {
    result[94]++;
    continue;
}
if (s.equals("should")) {
    result[95]++;
    continue;
}
if (s.equals("since")) {
    result[96]++;
    continue;
}
if (s.equals("so")) {
    result[97]++;
    continue;
}
if (s.equals("some")) {
    result[98]++;
    continue;
}
if (s.equals("somebody")) {
    result[99]++;
    continue;
}
if (s.equals("someone")) {
    result[100]++;
    continue;
}
if (s.equals("something")) {
    result[101]++;
    continue;
}
if (s.equals("such")) {
    result[102]++;
    continue;
}
if (s.equals("than")) {
    result[103]++;
    continue;
}
if (s.equals("that")) {
    result[104]++;
    continue;
}
if (s.equals("he")) {
    result[105]++;
    continue;
}
if (s.equals("their")) {
    result[106]++;
    continue;
}
if (s.equals("them")) {
    result[107]++;
    continue;
}
if (s.equals("these")) {
    result[108]++;
    continue;
}
if (s.equals("they")) {
    result[109]++;
    continue;
}
if (s.equals("this")) {
```

```
        result[110]++;
        continue;
    }
    if (s.equals("those")) {
        result[111]++;
        continue;
    }
    if (s.equals("though")) {
        result[112]++;
        continue;
    }
    if (s.equals("through")) {
        result[113]++;
        continue;
    }
    if (s.equals("till")) {
        result[114]++;
        continue;
    }
    if (s.equals("to")) {
        result[115]++;
        continue;
    }
    if (s.equals("toward")) {
        result[116]++;
        continue;
    }
    if (s.equals("towards")) {
        result[117]++;
        continue;
    }
    if (s.equals("under")) {
        result[118]++;
        continue;
    }
    if (s.equals("unless")) {
        result[119]++;
        continue;
    }
    if (s.equals("unlike")) {
        result[120]++;
        continue;
    }
    if (s.equals("until")) {
        result[121]++;
        continue;
    }
    if (s.equals("up")) {
        result[122]++;
        continue;
    }
    if (s.equals("upon")) {
        result[123]++;
        continue;
    }
    if (s.equals("upper")) {
        result[124]++;
        continue;
    }
    if (s.equals("us")) {
        result[125]++;
        continue;
    }
    if (s.equals("used")) {
        result[126]++;
        continue;
    }
    if (s.equals("via")) {
        result[127]++;
        continue;
```

```
}
if (s.equals("we")) {
    result[128]++;
    continue;
}
if (s.equals("what")) {
    result[129]++;
    continue;
}
if (s.equals("whatever")) {
    result[130]++;
    continue;
}
if (s.equals("when")) {
    result[131]++;
    continue;
}
if (s.equals("where")) {
    result[132]++;
    continue;
}
if (s.equals("whether")) {
    result[133]++;
    continue;
}
if (s.equals("which")) {
    result[134]++;
    continue;
}
if (s.equals("while")) {
    result[135]++;
    continue;
}
if (s.equals("who")) {
    result[136]++;
    continue;
}
if (s.equals("whoever")) {
    result[137]++;
    continue;
}
if (s.equals("whom")) {
    result[138]++;
    continue;
}
if (s.equals("whose")) {
    result[139]++;
    continue;
}
if (s.equals("will")) {
    result[140]++;
    continue;
}
if (s.equals("with")) {
    result[141]++;
    continue;
}
if (s.equals("within")) {
    result[142]++;
    continue;
}
if (s.equals("without")) {
    result[143]++;
    continue;
}
if (s.equals("worth")) {
    result[144]++;
    continue;
}
if (s.equals("would")) {
```

```
                result[145]++;
                continue;
            }
            if (s.equals("yes")) {
                result[146]++;
                continue;
            }
            if (s.equals("you")) {
                result[147]++;
                continue;
            }
            if (s.equals("your")) {
                result[148]++;
                continue;
            }
            if (s.equals("not")){
                result[149]++;
                continue;
            }
        }
        return result;
    }

     * This method returns misc statistical data about the input text.
     * @param text Input text to be analyzed.
     * @param words Array of words of input text.
     * @return Coefficients retrieved after analysis.
    private float[] getLenghtsCount(String text, String[] words){
        float[] res = new float[25];

        Map<Integer, Integer> count = new HashMap<Integer, Integer>();

        float total_chars = 0;
        for(int i = 0; i<=13; i++){
            count.put(i, 0);
        }
        for(String word: words){
            total_chars += word.length();
            if (word.length() < 13){
                count.put(word.length(), count.get(word.length())+1);
            } else {
                count.put(13, count.get(13)+1);
            }
        }

        int sentence_num = 0;
// split text into sentences
        BreakIterator iterator = BreakIterator.getSentenceInstance(Locale.US);
        iterator.setText(text);
        int start = iterator.first();
        for (int end = iterator.next();
             end != BreakIterator.DONE;
             start = end, end = iterator.next()) {
            sentence_num++; // count number of the sentences
            String current = text.substring(start,end);
            if(!      ((current.toLowerCase().charAt(0)      ==      'i'      &&
current.toLowerCase().charAt(1) == ' ') || (current.toLowerCase().charAt(0) ==
'w' && current.toLowerCase().charAt(1) == 'e' && current.toLowerCase().charAt(2)
== ' '))){
                res[24]++; // number of 'I' or 'we' sentence beginnings
            }
            if(Character.isLowerCase(current.charAt(0))){
                res[22]++; // number of missing uppercase in the beginning of
the sentence
            }
            if(Character.isLetter(current.charAt(current.length()-1))){
                res[23]++; // number of missing any punctuation sign in the end
of sentence.
            }
        }
```

```
res[0] = sentence_num; // number of sentences
        res[1] = words.length; // number of words
        res[2] += count.get(1) + count.get(2) + count.get(3); // number of short
(<4 chars) words
        res[3] = total_chars / words.length; // average word length
        res[4] = total_chars / (float) sentence_num; // average chars per
sentence
        res[5] = (float) words.length / (float) sentence_num; // average words
per sentence
        res[6] = count.get(1); // number of words of 1 character
        res[7] = count.get(2); // number of words of 2 characters
        res[8] = count.get(3); // number of words of 3 characters
        res[9] = count.get(4); // number of words of 4 characters
        res[10] = count.get(5); // number of words of 5 characters
        res[11] = count.get(6); // number of words of 6 characters
        res[12] = count.get(7); // number of words of 7 characters
        res[13] = count.get(8); // number of words of 8 characters
        res[14] = count.get(9); // number of words of 9 characters
        res[15] = count.get(10); // number of words of 10 characters
        res[16] = count.get(11); // number of words of 11 characters
        res[17] = count.get(12); // number of words of 12 characters
        res[18] = count.get(13); // number of words with more than 12 characters
// count number of happy smiles in the text with regular expressions
Pattern                                    sad                                =
Pattern.compile(  "(\uD83E\uDD11|\uD83D\uDE06|:\\)|:=\\)|\uD83D\uDE0D|\uD83D\uD
E09|\uD83D\uDE0A|;\\)|\uD83D\uDE00|\uD83D\uDE01|\uD83D\uDE02|\uD83E\uDD23|\uD8
3D\uDE03|\uD83D\uDE04|\uD83D\uDE05|\uD83D\uDE0E|\uD83D\uDE42|\uD83D\uDE0B|â˜°ï
¸?|\uD83D\uDE0C|\uD83D\uDE43)");
        Matcher sadMatcher = sad.matcher(text);
        while (sadMatcher.find()){
            res[20]++;
        }
        // count number of 'lol' in the text (with different upper-lower cases
combinations)
        Pattern lol = Pattern.compile("lol");
        Matcher lolMatcher = lol.matcher(text.toLowerCase());
        while (lolMatcher.find()){
            res[21]++;
        }
        return res;
    }
private float[] getEmojiCount(String text){

        String[] emojis = new String[]{
"\uD83D\uDE00", // ðŸ˜€ grinning face

"\uD83D\uDE01", // ðŸ˜? Grinning Face With Smiling Eyes

"\uD83D\uDE02", // ðŸ˜‚ Face With Tears of Joy

"\uD83D\uDE03", // ðŸ˜ƒ Smiling Face With Open Mouth

"\uD83D\uDC4D", // ðŸ‘? Thumbs Up

"\uD83D\uDE09", // ðŸ˜‰ Winking face

"\uD83E\uDD17", // ðŸ¤— Hugging face

"â˜°", // â˜°ï¸?Smiling face

"\uD83D\uDE06", // ðŸ˜† Smiling Face With Open Mouth & Closed Eyes

"\uD83D\uDE04", // ðŸ˜„ Smiling Face With Open Mouth & Smiling Eyes

"\uD83D\uDE1B", // ðŸ˜› Face With Stuck-Out Tongue
```

```
"\uD83D\uDE1C", // ðŸ˜œ Face With Stuck-Out Tongue & Winking Eye

"\uD83D\uDE1D", // ðŸ˜? Face With Stuck-Out Tongue & Closed Eyes

"\uD83D\uDE23", // ðŸ˜£ Persevering Face

"\uD83D\uDE25", // ðŸ˜¥ Disappointed but Relieved Face

"â˜¹", // â˜¹ Frowning Face

"\uD83D\uDE2D", // ðŸ˜- Loudly Crying Face

"\uD83D\uDE41", // ðŸ™? Slightly Frowning Face

"\uD83D\uDE27", // ðŸ˜§ Anguished Face

"\uD83D\uDE29", // ðŸ˜© Weary Face

"\uD83D\uDE35", // ðŸ˜µ Dizzy Face

"\uD83D\uDE20", // ðŸ˜  Angry Face

"\uD83D\uDE22", // ðŸ˜¢ Crying Face

"\uD83D\uDC93", // ðŸ’“ beating heart

"\uD83D\uDC8B", // ðŸ’‹ kiss

"\uD83D\uDE4F", // ðŸ™? pray

"\uD83D\uDD90", // ðŸ–? raised hands with fingers

"âœŒ", // âœŒï¸?

"\uD83D\uDC94", // ðŸ’” broken heart

"\uD83D\uDC91", // ðŸ’‘ couple with heart

"\uD83D\uDC4E", // ðŸ‘Ž thumbs down

"\uD83D\uDE0D", // ðŸ˜? heart eyes

"\uD83E\uDD23", // ðŸ¤£ rofl

"\uD83D\uDE44", // ðŸ™„ Face With Rolling Eyes

"\uD83D\uDE42", // ðŸ™‚ Slightly Smiling Face

"\uD83D\uDE2B", // ðŸ˜« Tired Face

"\uD83D\uDE2A", // ðŸ˜ª Sleepy face

"\uD83D\uDE08", // ðŸ˜ˆ Smiling Face With Horns

"\uD83D\uDE2C", // ðŸ˜¬ Grimacing Face

"\uD83D\uDE16", // ðŸ˜- Confounded Face
```

```
"\u2764\uFE0F", // â?¤ heart

"\uD83D\uDC4C", // ðŸ‘Œ ok hand

};
        float[] res = new float[emojis.length];

        for (int i = 0; i < emojis.length; i++){
            res[i] = 0;
        }

        for (int i = 0; i < emojis.length; i++){
            Pattern p = Pattern.compile(emojis[i]);
            Matcher matcher = p.matcher(text);
            while (matcher.find()){
                res[i] += 1;
            }
        }

        return res;
    }
```

# Appendix E -Data Pre-Processing

- Pre-processing Email

```
package messageanalyzer.DataExport;

import java.io.*;
import java.util.ArrayList;
import java.util.List;
import java.util.regex.Matcher;
import java.util.regex.Pattern;

public class EmailExport {

    private static String[] quotesPatterns = new String[]{

              "From: (\\w+\\s)+\\[mailto:\\w+@(\\w+\\.{0,1})+\\]",
                    "On \\d{1,2} \\w+ \\d{4}, at \\d{1,2}:\\d{1,2}",
                    "From: [\\w\\s]+ \\[.+\\]"
};

/**
     * @param filename A path or a name of the CSV file with e-mail
     * @return A list of texts of Emails
     */
    public static List<String> getMessages(String dirname) throws Exception{
        List<String> result = new ArrayList<>();

        try {
            File directory = new File(dirname);
            System.out.println(directory.getAbsolutePath());
            for (final File mailFile : directory.listFiles()) {
                if (!mailFile.isDirectory()) {
                    String next = parseFile(mailFile);
                    result.add(next);
                }
            }
        } catch (Exception ex){
            ex.printStackTrace();
        }
        return result;
    }
    private static String parseFile(File file) throws Exception{
        FileInputStream is = new FileInputStream(file);
        List<Pattern> quotePatternsList = new ArrayList<>();
        for (String regex : quotesPatterns){
            quotePatternsList.add(Pattern.compile(regex));
        }
BufferedReader reader = new BufferedReader(new InputStreamReader(is,
"Unicode"));
        StringBuilder line = new StringBuilder();
        String current;
        boolean content = false;
        while ((current = reader.readLine()) != null){
            if (content){
                for (Pattern p : quotePatternsList){
                    Matcher matcher = p.matcher(current);


                    if (matcher.find()){


//System.out.println(line.toString());
                        return line.toString();
                    }
                }
if(current.startsWith("From:")){
                    break;
```

```
                }
                else {
                    line.append(current).append("\n");
                }
}
            if (current.equals("")){
                content = true;
            }

        }
        //System.out.println(line.toString());
        return line.toString();
    }
}
```

- Pre-processing SMS

```
Package messageanalyzer.DataExport;

import com.opencsv.CSVReader;

import java.io.FileInputStream;
import java.io.IOException;
import java.io.InputStreamReader;
import java.util.ArrayList;
import java.util.List;
/*
 * Files must be created with
 * -
https://play.google.com/store/apps/details?id=com.hupaiwen.smsexport&hl=en
 * - https://itunes.apple.com/us/app/export-ur-sms-pro-free-save-your-
messages-texts/id1086448303?mt=8
 */
public class SMSExport {
    public static List<String> getMessages(String filename) throws Exception{
        List<String> result = new ArrayList<String>();

        try {
            CSVReader reader = new CSVReader(new InputStreamReader(new
FileInputStream(filename), "UTF8"));
            List<String[]> csvData = reader.readAll();
            int i = 0;
            int z = 0;
            for (String[] row : csvData){
                if(i == 0){
                    for(int x=0; x<25; x++){
                        if(row[x].equals("||__Text")){
                            z = x;
                        }
                    }
                    i = 2;
                    //System.out.println("xxxx"+z);
                }
                else if(row[z+3].equals("0")){
                try {
                  result.add(row[z]);
                   //System.out.println("w "+row[z]);
                } catch (Exception ex){
                }
                }
            }
        } catch (IOException ex){
            ex.printStackTrace();
        }
        return result;
    }
}
```

- Pre-processing Twitter

```
Package messageanalyzer.DataExport;

import twitter4j.Status;
import twitter4j.Twitter;
import twitter4j.TwitterException;
import twitter4j.TwitterFactory;

import java.util.ArrayList;
import java.util.List;

import twitter4j.*;

public class TwitterExport {
    /**
     * This method returns a list of all tweets of given user.
     *
     * @param userName Name of the account to parse
     * @return List of the tweets of the given user
     * @throws TwitterException In case of failure
     */
    public static List<String> getMessages(String userName) throws
TwitterException{
        List<String> result = new ArrayList<String>();
        Twitter twitter = new TwitterFactory().getInstance();
        Paging paging = new Paging(1, 500);
        try {
            List<Status> statuses;
            statuses = twitter.getUserTimeline(userName,paging);

            for (Status status : statuses) {
                result.add(status.getText());
            }

            statuses = twitter.getUserTimeline(userName,new Paging(2,
500));
            for (Status status : statuses) {
                result.add(status.getText());
            }

            statuses = twitter.getUserTimeline(userName,new Paging(3,
500));
            for (Status status : statuses) {
                result.add(status.getText());
            }
        } catch (TwitterException te) {
            throw te;
        }
        return result;
    }

}
```

- Pre-processing Facebook

```
package messageanalyzer.DataExport;

import facebook4j.*;
import facebook4j.auth.AccessToken;

import java.util.ArrayList;
import java.util.List;


public class FacebookExport {
    /**
     * @param accessToken A string with generated access token
     * @return A list of strings with facebook posts
     * @throws Exception in case of user-not found or network error,
or if token is invalid
     */
    public static List<String> getMessages(String accessToken) throws
Exception{
        List<String> result = new ArrayList<String>();

        Facebook facebook = new FacebookFactory().getInstance();
        AccessToken at = new AccessToken(accessToken);
        // Set access token.
        facebook.setOAuthAccessToken(at);
        ResponseList<Post> feed = facebook.getPosts(new
Reading().limit(600));

        for(Post p : feed){
            result.add(p.getMessage());
        }

        return result;
    }
}
```

# Appendix F - Dataset

| user | Facebook | Twitter | Email | SMS | Total | #platforms |
|---|---|---|---|---|---|---|
| 1 | 71 | 583 | 83 | 19,141 | 19,878 | 4 |
| 2 | 46 | 20 | 161 | 403 | 630 | 4 |
| 3 | 27 | 599 | 72 | 37 | 735 | 4 |
| 4 | 28 | 579 | 30 | 1,718 | 2,355 | 4 |
| 5 | 189 | 584 | 386 | 852 | 2,011 | 4 |
| 6 | 90 | 595 | 21 | 6,071 | 6,777 | 4 |
| 7 | 48 | 590 | 202 | 2,687 | 3,527 | 4 |
| 8 | 95 | 270 | 49 | 1,279 | 1,693 | 4 |
| 9 | 76 | 590 | 80 | 4,729 | 5,475 | 4 |
| 10 | 68 | 146 | 51 | 3,611 | 3,876 | 4 |
| 11 | 56 | 105 | 38 | 29,710 | 29,909 | 4 |
| 12 | 139 | 46 | 314 | 207 | 706 | 4 |
| 13 | 76 | 587 | 109 | 45 | 817 | 4 |
| 14 | 117 | 594 | 39 | 5,243 | 5,993 | 4 |
| 15 | 97 | 596 | 125 | 25 | 843 | 4 |
| 16 | 106 | 106 | 43 | 523 | 778 | 4 |
| 17 | 69 | 575 | 145 | 10,596 | 11,385 | 4 |
| 18 | 71 | 26 | 34 | 909 | 1,040 | 4 |
| 19 | 132 | 591 | 165 | 0 | 888 | 3 |
| 20 | 175 | 0 | 79 | 7,512 | 7,766 | 3 |
| 21 | 189 | 151 | 24 | 0 | 364 | 3 |
| 22 | 37 | 589 | 20 | 0 | 645 | 3 |
| 23 | 142 | 176 | 20 | 0 | 337 | 3 |
| 24 | 26 | 0 | 38 | 4,499 | 4,563 | 3 |
| 25 | 216 | 0 | 26 | 548 | 790 | 3 |
| 26 | 145 | 586 | 22 | 0 | 753 | 3 |
| 27 | 131 | 0 | 120 | 27 | 278 | 3 |
| 28 | 35 | 590 | 178 | 0 | 803 | 3 |
| 29 | 51 | 62 | 129 | 0 | 242 | 3 |
| 30 | 0 | 22 | 83 | 979 | 1,084 | 3 |
| 31 | 140 | 163 | 35 | 0 | 338 | 3 |
| 32 | 195 | 98 | 774 | 0 | 1,067 | 3 |
| 33 | 34 | 184 | 28 | 0 | 246 | 3 |
| 34 | 29 | 573 | 66 | 0 | 668 | 3 |
| 35 | 208 | 87 | 1,323 | 0 | 1,618 | 3 |
| 36 | 100 | 583 | 104 | 0 | 787 | 3 |
| 37 | 145 | 564 | 28 | 0 | 737 | 3 |
| 38 | 39 | 0 | 0 | 627 | 666 | 3 |
| 39 | 23 | 120 | 0 | 4,237 | 4,380 | 3 |
| 40 | 86 | 0 | 71 | 144 | 301 | 3 |
| 41 | 97 | 578 | 214 | 0 | 889 | 3 |
| 42 | 128 | 26 | 96 | 0 | 250 | 3 |
| 43 | 200 | 211 | 53 | 0 | 464 | 3 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 44 | 0 | 26 | 30 | 0 | 56 | 2 |
| 45 | 0 | 19 | 23 | 0 | 42 | 2 |
| 46 | 109 | 0 | 116 | 0 | 225 | 2 |
| 47 | 72 | 0 | 310 | 0 | 382 | 2 |
| 48 | 60 | 406 | 0 | 0 | 466 | 2 |
| 49 | 0 | 20 | 74 | 0 | 94 | 2 |
| 50 | 126 | 0 | 309 | 0 | 435 | 2 |
| **No of users** | 46 | 41 | 47 | 26 | | |
| **No of samples** | 4.539 | 13.616 | 6.540 | 106.359 | | |

# Appendix G- Top Population-Based Feature of All Platforms with its Code

| **Twitter** EER (20.16)% for top 30 features | **SMS** EER (7.97)% for top 100 features | **Facebook** EER (25)% for top 275 features | **Email** EER (13.11)% for top 100 features |
|---|---|---|---|
| **Features** | **Features** | **Features** | **Features** |
| 31 | 27 | 52 | 29 |
| 231 | 232 | 55 | 38 |
| 55 | 231 | 54 | 50 |
| 3 | 52 | 1 | 55 |
| 1 | 209 | 2 | 39 |
| 2 | 215 | 231 | 102 |
| 52 | 233 | 213 | 51 |
| 54 | 274 | 212 | 52 |
| 213 | 1 | 3 | 231 |
| 39 | 3 | 214 | 42 |
| 32 | 228 | 210 | 228 |
| 48 | 51 | 209 | 227 |
| 210 | 53 | 32 | 43 |
| 214 | 2 | 232 | 212 |
| 27 | 54 | 8 | 213 |
| 212 | 210 | 22 | 3 |
| 227 | 55 | 48 | 54 |
| 219 | 213 | 23 | 13 |
| 209 | 214 | 12 | 58 |
| 23 | 211 | 228 | 6 |
| 21 | 236 | 233 | 14 |
| 232 | 212 | 211 | 31 |
| 8 | 23 | 58 | 126 |
| 22 | 12 | 11 | 214 |
| 224 | 36 | 216 | 27 |
| 233 | 107 | 236 | 56 |
| 51 | 58 | 17 | 26 |
| 211 | 56 | 4 | 1 |
| 4 | 217 | 53 | 21 |
| 19 | 8 | 24 | 107 |
| 18 | 17 | 40 | 73 |
| 17 | 216 | 18 | 224 |
| 226 | 218 | 274 | 65 |
| 6 | 4 | 28 | 9 |
| 44 | 113 | 6 | 2 |
| 15 | 18 | 21 | 4 |
| 13 | 48 | 217 | 11 |
| 216 | 22 | 36 | 193 |
| 274 | 11 | 7 | 220 |
| 28 | 21 | 218 | 217 |
| 5 | 230 | 15 | 219 |
| 12 | 219 | 224 | 24 |
| 30 | 10 | 29 | 18 |
| 218 | 239 | 265 | 110 |
| 24 | 267 | 219 | 28 |
| 45 | 59 | 215 | 7 |
| 7 | 222 | 10 | 15 |
| 16 | 15 | 19 | 210 |
| 14 | 7 | 16 | 209 |

| | | | |
|---|---|---|---|
| 228 | 14 | 26 | 221 |
| 56 | 28 | 222 | 211 |
| 11 | 206 | 27 | 222 |
| 215 | 24 | 9 | 5 |
| 217 | 26 | 14 | 206 |
| 225 | 68 | 38 | 216 |
| 184 | 38 | 225 | 22 |
| 10 | 19 | 220 | 233 |
| 9 | 16 | 5 | 12 |
| 26 | 5 | 107 | 215 |
| 29 | 32 | 25 | 118 |
| 221 | 229 | 51 | 23 |
| 142 | 6 | 78 | 101 |
| 25 | 220 | 134 | 218 |
| 223 | 9 | 223 | 10 |
| 58 | 25 | 226 | 16 |
| 220 | 227 | 221 | 40 |
| 222 | 119 | 65 | 25 |
| 57 | 221 | 20 | 19 |
| 236 | 77 | 56 | 60 |
| 20 | 74 | 227 | 17 |
| 107 | 238 | 136 | 225 |
| 206 | 241 | 126 | 8 |
| 87 | 80 | 13 | 197 |
| 174 | 181 | 266 | 223 |
| 53 | 125 | 59 | 142 |
| 42 | 156 | 109 | 53 |
| 59 | 223 | 169 | 41 |
| 126 | 40 | 235 | 109 |
| 187 | 76 | 206 | 207 |
| 253 | 224 | 87 | 114 |
| 68 | 86 | 114 | 174 |
| 163 | 13 | 68 | 90 |
| 229 | 101 | 31 | 187 |
| 103 | 29 | 115 | 78 |
| 239 | 163 | 174 | 88 |
| 101 | 122 | 43 | 34 |
| 156 | 114 | 113 | 68 |
| 38 | 188 | 200 | 113 |
| 100 | 79 | 101 | 184 |
| 122 | 187 | 275 | 167 |
| 136 | 163 | 77 | 51 |

## Population based feature code

```
# -*- coding: utf-8 -*-
"""
Created on Tue Jan 19 09:17:07 2018
import pandas as pd
import numpy as np
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import GradientBoostingClassifier
from sklearn import svm
from sklearn.model_selection import train_test_split
from imblearn.datasets import make_imbalance
from sklearn.metrics import f1_score, precision_score, recall_score
  #loading the dataset
from sklearn import metrics
```

```
#######################################################
#######################################################
experiment = 'static'
platform_name = 'All samples'
classifiers = {'GB':0, 'RF':1, 'SVM':0, 'KNN':0}
metric = {'F1':0, 'EER':1}
   Importing the dataset
dataset = pd.read_csv('All_Platforms_Samples.csv')
dataset.fillna(0, inplace=True)
#dataset = dataset[dataset['user'] < 15]
train_test_ratio = [30]
top_features = [10, 20 ,30 ,50 ,100, 275]
#######################################################
#######################################################
for ratio in train_test_ratio:
for num_features in top_features:
#removing duplicates rows
 #dataset = dataset.drop_duplicates()
    def dataPrep(dataset):
  #replacing Nan values with 0
  X = dataset.loc[:, 'sample':].values
 y = dataset['user'].values
#scalling dataset
from sklearn.preprocessing import StandardScaler
  sc = StandardScaler()
 X  = sc.fit_transform(X)
  return X, y
  #getting the most importnat features
 def feature_imp_RandomForest(X, y):
 #abstracting the classifier
  rf = RandomForestClassifier(n_estimators=300, max_depth=8, min_samples_leaf=4, max_fe
atures=0.2, n_jobs=-1, random_state=12345)
rf.fit(X, y)
  #returning the most k important features sorted by their importance
 selected_features =  X[:,rf.feature_importances_.argsort()[::-1][:num_features]]
   return selected_features
 #splitiing train/test sets
 def data_split(X, y_copy):
gen_num_samples = sum(y_copy[y_copy == 1])
  #undersampling the dataset by taking all geniun samples verses 10% of imposters sampl
es
  #Sampling here
  #make_imbalance is a libariry that compatable with sklearn th perform undersampling a
nd more
X, y_copy = make_imbalance(X, y_copy, ratio={0: gen_num_samples, 1: gen_num_samples}, r
andom_state=12345)
 # use only the most k importnat features
X_train, X_test, y_train, y_test = train_test_split(X, y_copy, test_size = ratio/100, r
andom_state=12345)   #perform over-sampling using SMOTE
   #sm = SMOTE(ratio='minority', random_state=42)
            #X_train, y_train = sm.fit_sample(X_train, y_train)
                       return X_train, y_train, X_test, pd.Series(y_test)
                    def KNN(X_train, y_train, X_test, y_test):
            #abstracting the classifier
            classifier = KNeighborsClassifier(n_neighbors=3, random_state=12345)
            classifier.fit(X_train, y_train)
            if metric['F1'] == 1:
                y_pred = classifier.predict(X_test)
                Acc = round(metrics.accuracy_score(y_pred,y_test),2)
                print('ACC', Acc)
                return y_pred, Acc
            if metric['EER'] == 1:
                y_pred = classifier.predict_proba(X_test)
                return y_pred
                           def RF(X_train, y_train, X_test, y_test):
        #abstracting the classifier
            classifier = RandomForestClassifier(random_state=12345)
            classifier.fit(X_train, y_train)
```

```
            if metric['F1'] == 1:
                y_pred = classifier.predict(X_test)
                Acc = round(metrics.accuracy_score(y_pred,y_test),2)
                print('ACC', Acc)
                return y_pred, Acc

            if metric['EER'] == 1:
                y_pred = classifier.predict_proba(X_test)
                return y_pred
      def SVM(X_train, y_train, X_test, y_test):
            #abstracting the classifier
          classifier = svm.SVC(probability=True, random_state=12345)
         classifier.fit(X_train, y_train)
                     if metric['F1'] == 1:
              y_pred = classifier.predict(X_test)
                Acc = round(metrics.accuracy_score(y_pred,y_test),2)
                print('KNN ACC', Acc)
               return y_pred, Acc
                        if metric['EER'] == 1:
                y_pred = classifier.predict_proba(X_test)
                return y_pred
        def GB(X_train, y_train, X_test, y_test):
                         #abstracting the classifier
            classifier = GradientBoostingClassifier(verbose=0, n_estimators=300, random
_state=12345)
           classifier.fit(X_train, y_train)
          if metric['F1'] == 1:
               y_pred = classifier.predict(X_test)
               Acc = round(metrics.accuracy_score(y_pred,y_test),2)
                print('GB ACC', Acc)
               return y_pred, Acc
                         if metric['EER'] == 1:
                y_pred = classifier.predict_proba(X_test)
               return y_pred
  #calculating f1
       def F_score(preds_org):
            recall = precision_score(y_test, preds_org)
           precision = recall_score(y_test, preds_org)
            f1 = f1_score(y_test, preds_org)
            print('F1: ', (round(f1,2)))
            ALL_Threshold_results['Recall'].append(round(f1,2))
            ALL_Threshold_results['Precision'].append(round(f1,2))
         ALL_Threshold_results['F1'].append(round(f1,2))
                       return f1, recall, precision

  #calculating EER
        def EER(preds_org):

 #these list are used to recored rates for each user
      FRR_list = []
            FAR_list = []
            EER_list= []
           Precision = 100
            #threshold loop
           for threshold in range(1, Precision):
               threshold = threshold/Precision
                #coping the predictions probabilites
                preds = preds_org.copy()

                #setting the class according to the threshold
                preds[preds >= threshold] = 1
               preds[preds < threshold] = 0

  #    converting numpy 2d array into 1d pandas series with class name with highest

   #    probability and returns single column
                preds = (pd.DataFrame(preds)).idxmax(axis=1)
```

```
  #getting the list of genuin samples indices
              gen_test_samples_index_list =  (y_test[y_test == 1]).index.tolist()

  #getting the list of imposters samples indices
  imp_test_samples_index_list =  (y_test[y_test == 0]).index.tolist()
               #getting the number of genuin samples
 num_gen_samples = (len(gen_test_samples_index_list))
#FRR is equal to the sum of rejected genuin samples divided by totoal number of genuin
samples
 FRR = ((num_gen_samples - (sum(preds.loc[gen_test_samples_index_list])))/num_gen_sampl
es)*100
     #getting the number of imposters samples
    num_imp_samples = (len(imp_test_samples_index_list))
     #FAR is equal to the sum of accepted imposters samples divided by totoal number of
 imposters
samples
    FAR =  (sum(preds.loc[imp_test_samples_index_list])/num_imp_samples)*100
#appending both FRR and FAR into their lists to be used later in the next for loop to c
alculate EER
 FRR_list.append(FRR)
    FAR_list.append(FAR)
  #(optional) deleting predction seires to make sure each class (subject) gets new one
out of the classifier
 del preds
  for EER in zip(FRR_list, FAR_list):
   #appending the absolute difference between FRR and FAR
  EER_list.append(abs(EER[0]-EER[1]))
  #getting the index of the minimum value of absolute difference between FRR and FAR
   min_diff_value_index = EER_list.index(min(EER_list))
 #EER is equal to the(FRR + FAR)/2 for the minimum value of absolute difference between
 FRR and FAR
 min_EER = (FRR_list[min_diff_value_index] + FAR_list[min_diff_value_index])/2
 #printing the prediction result for each class iteration
  print('Threshold: ',(min_diff_value_index+1)/Precision)
            print('FRR: ',FRR_list[min_diff_value_index])
            print('FAR: ',FAR_list[min_diff_value_index])
            print('EER: ',min_EER)

            #appending the obtained result into the reuslt dict
            ALL_Threshold_results['Threshold'].append((min_diff_value_index+1)/Precisio
n)
            ALL_Threshold_results['FAR'].append(FAR_list[min_diff_value_index])
            ALL_Threshold_results['FRR'].append(FRR_list[min_diff_value_index])
            ALL_Threshold_results['EER'].append(min_EER)
            return min_EER
                 X, y = dataPrep(dataset)
                  X = feature_imp_RandomForest(X, y)
        #X = dataset.iloc[:,1:]
        #y = dataset.iloc[:,0]
        if metric['F1'] == 1:
            #summing all EER to be divided by the number of users
           f1_counter = 0
          recall_counter = 0
            precision_counter = 0
            Acc_counter = 0
            user_counter = 0
            #dict to store results
           ALL_Threshold_results = {'User':[],'Acc':[],'F1':[],'Recall':[],'Precision':
[]}
        if metric['EER'] == 1:
            #summing all EER to be divided by the number of users
            EER_counter = 0
            FAR_counter = 0
            FRR_counter = 0
            Acc_counter = 0
            user_counter = 0
            #dict to store results
```

```
            ALL_Threshold_results = {'User':[],'EER':[],'FAR':[],'FRR':[],'Threshold':[
]}
        #iterating over users (subject column)
        for subject in np.unique(y):#getting the unique vaues of y to make sure that mi
ssing classes don't break the loop
            #user counter to be used for division the result
            user_counter +=1
            #appending user number to the result dict
         ALL_Threshold_results['User'].append(subject)
  print(subject)

            #coping y series  (calss) at each iteration so the subsequent modification
does not change the orginal series
            y_copy = y.copy()

            #setting calss lable according to the current iteration
            y_copy[y_copy != subject] = 0
            y_copy[y_copy == subject] = 1

          #spliting the dataset into train/test using data_split() function
            X_train, y_train, X_test, y_test = data_split(X, y_copy)
         if classifiers['GB'] == 1:
               if metric['F1'] == 1:
                    preds_org, Acc = GB(X_train, y_train, X_test,y_test)
                                if metric['EER'] == 1:
                    preds_org = GB(X_train, y_train, X_test,y_test)

                filename = experiment+'_'+platform_name+'_GB_'+str(ratio)+'_'+str(num_f
eatures)
                                if classifiers['RF'] == 1:
               if metric['F1'] == 1:
                    preds_org, Acc = RF(X_train, y_train, X_test,y_test)
                                if metric['EER'] == 1:
                    preds_org = RF(X_train, y_train, X_test,y_test)

                filename = experiment+'_'+platform_name+'_RF_'+str(ratio)+'_'+str(num_f
eatures)
            if classifiers['SVM'] == 1:
                if metric['F1'] == 1:
                  preds_org, Acc = SVM(X_train, y_train, X_test,y_test)
                   if metric['EER'] == 1:
                    preds_org = SVM(X_train, y_train, X_test,y_test)

                filename = experiment+'_'+platform_name+'_SVM_'+str(ratio)+'_'+str(num_
features)                               if classifiers['KNN'] == 1:
               if metric['F1'] == 1:
                   preds_org, Acc = KNN(X_train, y_train, X_test,y_test)
                               if metric['EER'] == 1:
                    preds_org = KNN(X_train, y_train, X_test,y_test)

 filename = experiment+'_'+platform_name+'_KNN_'+str(ratio)+'_'+str(num_features)
           if metric['F1'] == 1:
                ALL_Threshold_results['Acc'].append(round(Acc,2))
                          if metric['F1'] == 1:
              f1, recall, precision = F_score(preds_org)
           if metric['EER'] == 1:
              EER_ = EER(preds_org)
            #summing all EER to be divided by the number of users
   if metric['F1'] == 1:
       Acc_counter += Acc
       recall_counter += recall
       precision_counter += precision
       f1_counter += f1
                  if metric['EER'] == 1:
       Acc_counter += Acc
       FAR_counter += FAR
       FRR_counter += FRR
EER_counter += EER_
```

```
   if metric['F1'] == 1:
                print('Avareged F1: ',round(f1_counter/user_counter,2))
            if metric['EER'] == 1:
                print('Avareged EER: ',round(EER_counter/user_counter,2))
              #converting 'result' dict into dataframe to write the result into excel f
ile
   Result = pd.DataFrame.from_dict(ALL_Threshold_results)
       if metric['F1'] == 1:
           #appending EER/usernumber into result dataframe
            Result['Avg_F1'] = pd.Series(round(f1_counter / len(np.unique(y)), 2))
           Result['Avg_recall'] = pd.Series(round(recall_counter / len(np.unique(y)), 2
))                Result['Avg_precision'] = pd.Series(round(precision_counter / len(np.
unique(y)), 2))
   if metric['EER'] == 1:
          #appending EER/usernumber into result dataframe
            Result['Avg_EER'] = pd.Series(round(EER_counter / len(np.unique(y)), 2))

           Result['Avg_FAR'] = pd.Series(round(Result['FAR'].mean(), 2))
            Result['Avg_FRR'] = pd.Series(round(Result['FRR'].mean(), 2))
                                        if metric['F1'] == 1:
            Result['Avg_Acc'] = pd.Series(round(Acc_counter / len(np.unique(y)), 2))
                        #seting user col to be the index
       Result = Result.set_index('User')
        #wrtiting the result dataframe into a file
  if metric['F1'] == 1:
      Result.to_excel('results/'+experiment+'/'+platform_name+'/'+'F1_'+filename+'.xlsx
')             if metric['EER'] == 1:
            Result.to_excel('results/'+experiment+'/'+platform_name+'/'+'EER_'+filename
+'.xlsx')
```

# Appendix H –Indvidual Features for Authors Across Platforms

**User 1**

| 1 | 2 | 3 | 4 |
|---|---|---|---|
| Twitter | SMS | FB | Email |
| 55 | 28 | 1 | 1 |
| 214 | 1 | 55 | 224 |
| 53 | 233 | 13 | 213 |
| 215 | 232 | 9 | 52 |
| 1 | 275 | 210 | 49 |
| 40 | 2 | 214 | 53 |
| 2 | 210 | 215 | 229 |
| 3 | 234 | 2 | 44 |
| 58 | 229 | 213 | 43 |
| 227 | 55 | 16 | 40 |
| 4 | 53 | 3 | 18 |
| 23 | 3 | 22 | 12 |
| 32 | 4 | 234 | 228 |
| 210 | 211 | 212 | 220 |
| 213 | 215 | 5 | 15 |
| 29 | 214 | 23 | 214 |
| 22 | 59 | 17 | 16 |
| 234 | 24 | 12 | 234 |
| 211 | 212 | 24 | 4 |
| 19 | 213 | 4 | 218 |
| 18 | 216 | 211 | 212 |
| 5 | 56 | 102 | 115 |
| 9 | 54 | 29 | 23 |
| 220 | 52 | 218 | 2 |
| 56 | 108 | 19 | 19 |
| 24 | 219 | 53 | 25 |
| 16 | 218 | 217 | 217 |

**User 15**

| 1 | 2 | 3 | 4 |
|---|---|---|---|
| Twitter | SMS | FB | Email |
| 1 | 1 | 1 | 1 |
| 32 | 28 | 213 | 40 |
| 2 | 3 | 224 | 20 |
| 33 | 214 | 275 | 213 |
| 55 | 213 | 9 | 227 |
| 3 | 220 | 22 | 30 |
| 24 | 2 | 214 | 228 |
| 4 | 4 | 3 | 103 |
| 19 | 24 | 2 | 59 |
| 224 | 211 | 219 | 214 |
| 20 | 12 | 29 | 49 |
| 56 | 8 | 33 | 4 |
| 214 | 233 | 12 | 2 |
| 213 | 27 | 211 | 12 |
| 7 | 59 | 5 | 223 |
| 15 | 6 | 19 | |
| 215 | 212 | 215 | |
| 40 | | 25 | |
| 5 | | 8 | |
| 211 | | 4 | |
| 218 | | 18 | |
| 9 | | 24 | |
| 212 | | 221 | |
| 13 | | 7 | |
| 16 | | 212 | |
| | | 28 | |
| | | 16 | |

**User 18**

| 1 | 2 | 3 | 4 |
|---|---|---|---|
| Twitter | SMS | FB | Email |
| 1 | 1 | 1 | 1 |
| 32 | 52 | 233 | 44 |
| 2 | 28 | 55 | 49 |
| 3 | 233 | 4 | 43 |
| 213 | 211 | 229 | 210 |
| 211 | 2 | 213 | 56 |
| 22 | 215 | 215 | 29 |
| 214 | 3 | 218 | 22 |
| 23 | 212 | 214 | 234 |
| 4 | 237 | 216 | 53 |
| 7 | 12 | 9 | 8 |
| 5 | 55 | 212 | 223 |
| 13 | | 3 | 6 |
| 19 | | 2 | 102 |
| 16 | | 20 | 214 |
| 9 | | 24 | 26 |
| 25 | | 13 | 15 |
| 56 | | 10 | 215 |
| 218 | | 25 | 10 |
| | | 211 | 17 |
| | | 18 | |
| | | 219 | |
| | | 8 | |
| | | 29 | |
| | | 234 | |
| | | 23 | |
| | | 22 | |

**User 21**

Platforms

| 1 | 2 | 3 | 4 |
|---|---|---|---|
| Twitter | SMS | FB | Email |
| 1 | N/A | 1 | 1 |
| 214 | N/A | 211 | 213 |
| 215 | N/A | 212 | 13 |
| 3 | N/A | 215 | 215 |
| 217 | N/A | 9 | 25 |
| 2 | N/A | 2 | 214 |
| 212 | N/A | 3 | 223 |
| 211 | N/A | 214 | 4 |
| 33 | N/A | 19 | 12 |
| 32 | N/A | 53 | 10 |
| 9 | N/A | 213 | 217 |
| 53 | N/A | 4 | 23 |
| 232 | N/A | 217 | 212 |
| 4 | N/A | 33 | 216 |
| 19 | N/A | 13 | 15 |
| 213 | N/A | 22 | 7 |
| 24 | N/A | 18 | 19 |

**User25**

Platforms

| 1 | 2 | 3 | 4 |
|---|---|---|---|
| Twitter | SMS | FB | Email |
| N/A | 1 | 1 | 1 |
| N/A | 28 | 215 | 108 |
| N/A | 22 | 214 | 207 |
| N/A | 214 | 55 | 223 |
| N/A | 213 | 3 | 15 |
| N/A | 215 | 2 | 3 |
| N/A | 217 | 13 | 29 |
| N/A | 3 | 4 | 10 |
| N/A | 211 | 213 | 226 |
| N/A | 2 | 11 | 2 |
| N/A | 56 | 18 | 213 |
| N/A | | 211 | 212 |
| N/A | | 24 | 211 |
| N/A | | 53 | 5 |
| N/A | | 219 | 69 |
| N/A | | 19 | 27 |
| N/A | | 217 | 9 |

**User 30**

Platforms

| 1 | 2 | 3 | 4 |
|---|---|---|---|
| Twitter | SMS | FB | Email |
| 1 | 1 | N/A | 1 |
| 53 | 53 | N/A | 213 |
| 17 | 233 | N/A | 40 |
| 213 | 28 | N/A | 16 |
| 3 | 215 | N/A | 7 |
| 12 | 210 | N/A | 214 |
| 24 | 54 | N/A | 30 |
| 2 | 211 | N/A | 39 |
| 4 | 234 | N/A | 59 |
| 220 | 214 | N/A | 6 |
| 59 | 3 | N/A | 215 |
| 23 | 213 | N/A | 221 |
| 19 | 55 | N/A | 66 |
| 20 | 2 | N/A | 49 |
| 8 | 229 | N/A | 234 |
| 22 | 22 | N/A | 175 |
| 9 | 212 | N/A | 127 |

# Appendix I - Statistical Process for Word Count

## 1. Twitter

| Users | Total messages | Total words among all samples | Average words per user |
|---|---|---|---|
| 1 | 583 | 6225 | 10.7 |
| 2 | 20 | 364 | 18.2 |
| 3 | 599 | 9386 | 15.7 |
| 4 | 579 | 6031 | 10.4 |
| 5 | 584 | 6662 | 11.4 |
| 6 | 595 | 7056 | 11.9 |
| 7 | 590 | 8553 | 14.5 |
| 8 | 270 | 4348 | 16.1 |
| 9 | 590 | 5736 | 9.7 |
| 10 | 146 | 1398 | 9.6 |
| 11 | 105 | 1083 | 10.3 |
| 12 | 46 | 688 | 15.0 |
| 13 | 587 | 7103 | 12.1 |
| 14 | 594 | 8264 | 13.9 |
| 15 | 596 | 9808 | 16.5 |
| 16 | 106 | 876 | 8.3 |
| 17 | 575 | 8492 | 14.8 |
| 18 | 26 | 167 | 6.4 |
| 19 | 591 | 11426 | 19.3 |
| 20 | 0 | 0 | 0 |
| 21 | 151 | 946 | 6.3 |
| 22 | 589 | 6918 | 11.7 |
| 23 | 176 | 2420 | 13.8 |
| 24 | 0 | 0 | 0 |
| 25 | 0 | 0 | 0 |
| 26 | 586 | 7315 | 12.5 |
| 27 | 0 | 0 | 0 |
| 28 | 590 | 8057 | 13.7 |
| 29 | 62 | 1136 | 18.3 |
| 30 | 22 | 381 | 17.3 |
| 31 | 163 | 1960 | 12.0 |
| 32 | 98 | 1567 | 16.0 |
| 33 | 184 | 2069 | 11.2 |
| 34 | 573 | 7647 | 13.3 |
| 35 | 87 | 1439 | 16.5 |
| 36 | 583 | 7632 | 13.1 |
| 37 | 564 | 6779 | 12.0 |
| 38 | 0 | 0 | 0 |
| 39 | 120 | 1343 | 11.2 |
| 40 | 0 | 0 | 0 |
| 41 | 578 | 6897 | 11.9 |
| 42 | 26 | 341 | 13.1 |
| 43 | 211 | 2825 | 13.4 |
| 44 | 26 | 310 | 11.9 |
| 45 | 20 | 233 | 11.7 |
| 46 | 0 | 0 | 0 |
| 47 | 0 | 0 | 0 |
| 48 | 406 | 4658 | 11.5 |
| 49 | 20 | 333 | 16.7 |
| 50 | 0 | 0 | 0 |
| **Total** | 13,617 | 176,872 | |
| **1th median** | | | 12.5 |
| **2th Median** | | | 11.4 |
| **#users** | 41 | | |

***Average words per message=** Total words among all samples/Total messages.

***Calculating Median**

1- First median: 12.5 Approximately13

2- Sorting data from smallest to largest.

3- Second median: 11.4 Approximately 11

## 2. SMS

| Users | Total messages | Total words among all samples | Average words per user |
|---|---|---|---|
| 1 | 19,141 | 175427 | 9.2 |
| 2 | 403 | 3884 | 9.6 |
| 3 | 37 | 285 | 7.7 |
| 4 | 1,718 | 18175 | 10.6 |
| 5 | 852 | 5883 | 6.9 |
| 6 | 6,071 | 39271 | 6.5 |
| 7 | 2,687 | 16215 | 6.0 |
| 8 | 1,279 | 20769 | 16.2 |
| 9 | 4,729 | 41876 | 8.9 |
| 10 | 3,611 | 40701 | 11.3 |
| 11 | 29,710 | 221552 | 7.5 |
| 12 | 207 | 3387 | 16.4 |
| 13 | 45 | 629 | 14.0 |
| 14 | 5,243 | 50380 | 9.6 |
| 15 | 25 | 347 | 13.9 |
| 16 | 523 | 3209 | 6.1 |
| 17 | 10,596 | 63334 | 6.0 |
| 18 | 909 | 15337 | 16.9 |
| 19 | 0 | 0 | 0 |
| 20 | 7,512 | 211650 | 28.2 |
| 21 | 0 | 0 | 0 |
| 22 | 0 | 0 | 0 |
| 23 | 0 | 0 | 0 |
| 24 | 4,499 | 31440 | 7.0 |
| 25 | 548 | 8335 | 15.2 |
| 26 | 0 | 0 | 0 |
| 27 | 27 | 800 | 29.6 |
| 28 | 0 | 0 | 0 |
| 29 | 0 | 0 | 0 |
| 30 | 979 | 12998 | 13.3 |
| 31 | 0 | 0 | 0 |
| 32 | 0 | 0 | 0 |
| 33 | 0 | 0 | 0 |
| 34 | 0 | 0 | 0 |
| 35 | 0 | 0 | 0 |
| 36 | 0 | 0 | 0 |
| 37 | 0 | 0 | 0 |
| 38 | 627 | 4460 | 7.1 |
| 39 | 4,237 | 49006 | 11.6 |
| 40 | 144 | 2473 | 17.2 |
| 41 | 0 | 0 | 0 |
| 42 | 0 | 0 | 0 |
| 43 | 0 | 0 | 0 |
| 44 | 0 | 0 | 0 |
| 45 | 0 | 0 | 0 |
| 46 | 0 | 0 | 0 |
| 47 | 0 | 0 | 0 |
| 48 | 0 | 0 | 0 |
| 49 | 0 | 0 | 0 |
| 50 | 0 | 0 | 0 |
| **Total** | 106,359 | | |
| **1th Median** | | | 10 |
| **2th Median** | | | 7 |
| **#users** | 26 | | |

***Average words per message=** Total words among all samples/Total messages.

***Calculating Median**

1- First median: 10.1 Approximately 10

2- Sorting data from smallest to largest.

3- Second median: 7.1 Approximately 7

## 3- Facebook

| User | Total messages | Total words among all samples | Average words per user |
|---|---|---|---|
| 1 | 71 | 905 | 12.7 |
| 2 | 46 | 315 | 6.8 |
| 3 | 27 | 359 | 13.3 |
| 4 | 28 | 310 | 11.1 |
| 5 | 189 | 1526 | 8.1 |
| 6 | 90 | 2052 | 22.8 |
| 7 | 48 | 279 | 5.8 |
| 8 | 95 | 897 | 9.4 |
| 9 | 76 | 586 | 7.7 |
| 10 | 68 | 488 | 7.2 |
| 11 | 56 | 406 | 7.3 |
| 12 | 139 | 2416 | 17.4 |
| 13 | 76 | 661 | 8.7 |
| 14 | 117 | 2049 | 17.5 |
| 15 | 97 | 1305 | 13.5 |
| 16 | 106 | 592 | 5.6 |
| 17 | 69 | 501 | 7.3 |
| 18 | 71 | 702 | 9.9 |
| 19 | 132 | 14260 | 108.0 |
| 20 | 175 | 2627 | 15.0 |
| 21 | 189 | 750 | 4.0 |
| 22 | 37 | 361 | 9.8 |
| 23 | 142 | 1664 | 11.7 |
| 24 | 26 | 243 | 9.3 |
| 25 | 216 | 4574 | 21.2 |
| 26 | 145 | 796 | 5.5 |
| 27 | 131 | 2053 | 15.7 |
| 28 | 35 | 217 | 6.2 |
| 29 | 51 | 615 | 12.1 |
| 30 | 0 | 0 | 0 |
| 31 | 140 | 1328 | 9.5 |
| 32 | 195 | 4328 | 22.2 |
| 33 | 34 | 250 | 7.4 |
| 34 | 29 | 274 | 9.4 |
| 35 | 208 | 2775 | 13.3 |
| 36 | 100 | 1007 | 10.1 |
| 37 | 145 | 1375 | 9.5 |
| 38 | 39 | 578 | 14.8 |
| 39 | 23 | 111 | 4.8 |
| 40 | 86 | 865 | 10.1 |
| 41 | 97 | 1570 | 16.2 |
| 42 | 128 | 1117 | 8.7 |
| 43 | 200 | 3113 | 15.6 |
| 44 | 0 | 0 | 0 |
| 45 | 0 | 0 | 0 |
| 46 | 109 | 1271 | 11.7 |
| 47 | 72 | 1261 | 17.5 |
| 48 | 60 | 537 | 9.0 |
| 49 | 0 | 0 | 0 |
| 50 | 126 | 1578 | 12.5 |
| Total | 4,539 | | |
| 1th Median | | | 10 |
| 2th Median | | | 7.7 |
| #users | 46 | | |

***Calculating Median**

1- First median: 10

2- Sorting data from smallest to largest.

3- Second median: 7.7 approximately

## 4- Email

| Users | Total messages | Total words among all samples | Average words per message |
|---|---|---|---|
| 1 | 83 | 3861 | 46.5 |
| 2 | 161 | 6418 | 39.9 |
| 3 | 72 | 5262 | 73.1 |
| 4 | 30 | 1254 | 41.8 |
| 5 | 386 | 18743 | 48.6 |
| 6 | 21 | 828 | 39.4 |
| 7 | 202 | 8226 | 40.7 |
| 8 | 49 | 3001 | 61.2 |
| 9 | 80 | 4000 | 50.0 |
| 10 | 51 | 2341 | 45.9 |
| 11 | 38 | 2230 | 58.7 |
| 12 | 314 | 17745 | 56.5 |
| 13 | 109 | 5454 | 50.0 |
| 14 | 39 | 2525 | 64.7 |
| 15 | 125 | 9261 | 74.1 |
| 16 | 43 | 2240 | 52.1 |
| 17 | 145 | 5541 | 38.2 |
| 18 | 34 | 1463 | 43.0 |
| 19 | 165 | 11928 | 72.3 |
| 20 | 79 | 3675 | 46.5 |
| 21 | 24 | 316 | 13.2 |
| 22 | 20 | 610 | 30.5 |
| 23 | 20 | 614 | 30.7 |
| 24 | 38 | 1664 | 43.8 |
| 25 | 26 | 1067 | 41.0 |
| 26 | 22 | 747 | 34.0 |
| 27 | 120 | 11888 | 99.1 |
| 28 | 178 | 15705 | 88.2 |
| 29 | 129 | 7416 | 57.5 |

| User | Total messages | Total words among all samples | Average words per message |
|---|---|---|---|
| 30 | 83 | 4668 | 56.2 |
| 31 | 35 | 1852 | 52.9 |
| 32 | 774 | 37401 | 48.3 |
| 33 | 28 | 2199 | 78.5 |
| 34 | 66 | 3690 | 55.9 |
| 35 | 1,323 | 206561 | 156.1 |
| 36 | 104 | 3714 | 35.7 |
| 37 | 28 | 2247 | 80.3 |
| 38 | 0 | 0 | 0 |
| 39 | 0 | 0 | 0 |
| 40 | 71 | 4390 | 61.8 |
| 41 | 214 | 18605 | 86.9 |
| 42 | 96 | 7871 | 82.0 |
| 43 | 53 | 3219 | 60.7 |
| 44 | 30 | 1308 | 43.6 |
| 45 | 23 | 1326 | 57.7 |
| 46 | 116 | 5256 | 45.3 |
| 47 | 310 | 10708 | 34.5 |
| 48 | 0 | 0 | 0 |
| 49 | 74 | 2410 | 32.6 |
| 50 | 309 | 10321 | 33.4 |
| **Total** | 6,540 | | |
| **1th Median** | 139 | | 50 |
| **2th Median** | 72 | | 41 |
| **#users** | 47 | | |

***Average words per message=** Total words among all samples/Total messages.

***Calculating Median**

1- First median: 50

2- Sorting data from smallest to largest.

3- Second median: 40.85 Approximately 41

# Appendix J -Data collection procedures for each platform

The scenario for the data collection procedure varies from one platform to another. The following explains in detail the procedures of how the data were exported and extracted from each platform:

- **SMS platform**

1- A software tool called Jihosoft Phone Transfer has been used to export SMS text messages from the user's mobile phone to desktop; this software tool is available online at: http://www.jihosoft.com/mobile/phone-transfer.html. It can work with iOS and Android) mobile phones, as shown below in Figure 4-5 and Figure 4-6.


Figure 4-5: Jihosoft Phone Transfer main interface

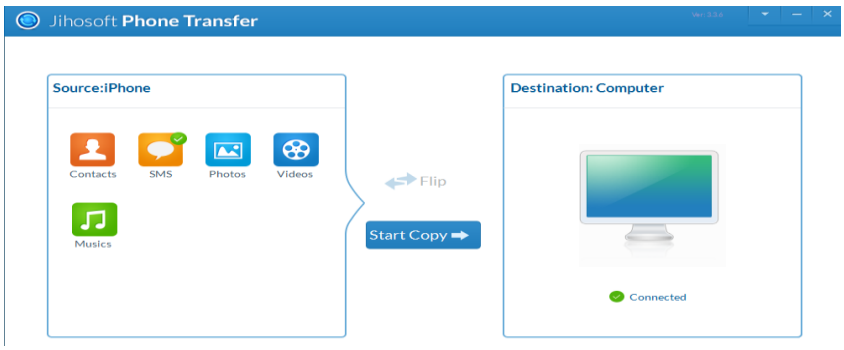


Figure 4-6: Selecting data type window

2- The above software was used to export SMS text message samples from the user and then saved as a Jscript script file on the desktop, as shown in Figure 4-7 below:


Figure 4-7: Exported SMS text

3- A software tool called JSON-CSV was used to convert the Jscript Script files to (.CSV) files. This software is available online at https://json-csv.en.softonic.com/ (see Figure 4-8 below):


Figure 4-8: JSON-CSV converter icon

4- The CSV file was stored in (.CSV) format in an SMS Output file on the desktop.

5- The message data parser, as shown below in, imported the CSV SMS output file to calculate the features.

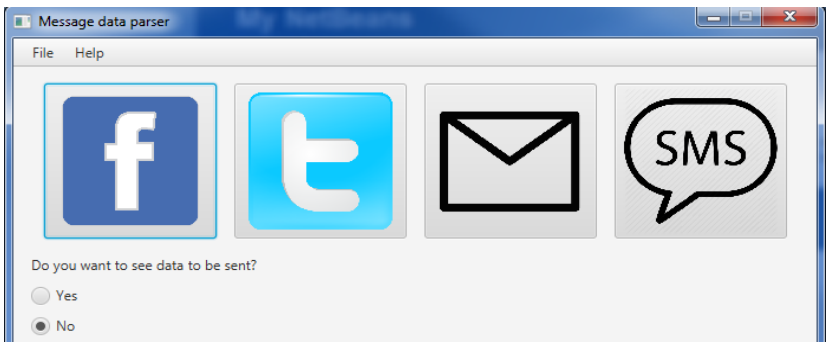


Figure 4-9: Message data parser main interface

6- The features were calculated from the text message, and contained only numbers. The Jscript Script file and CSV containing the SMS text output file were deleted from the desktop and hard drive before the user left; the only data that was obtained was the calculated features, as shown in below Figure 4-10 below:

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | F9 | F10 | F11 | F12 | F13 | F14 | F15 | F16 |
| 2 | 36 | 20 | 19 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 17 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 45 | 17 | 17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 11 | 10 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 2 | 0 | 0 |
| 6 | 69 | 68 | 1 | 5 | 0 | 3 | 1 | 4 | 3 | 1 | 2 | 6 | 0 | 1 | 1 | 4 |
| 7 | 105 | 103 | 5 | 6 | 0 | 5 | 3 | 12 | 1 | 1 | 5 | 10 | 0 | 2 | 2 | 3 |
| 8 | 37 | 35 | 2 | 3 | 2 | 1 | 3 | 4 | 0 | 0 | 2 | 1 | 0 | 1 | 0 | 2 |
| 9 | 73 | 71 | 3 | 4 | 3 | 1 | 1 | 9 | 0 | 1 | 5 | 4 | 0 | 3 | 2 | 3 |
| 10 | 79 | 78 | 1 | 8 | 5 | 5 | 4 | 8 | 0 | 1 | 3 | 4 | 0 | 0 | 2 | 6 |
| 11 | 25 | 24 | 1 | 3 | 2 | 1 | 3 | 3 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 1 |
| 12 | 33 | 32 | 2 | 3 | 0 | 0 | 3 | 2 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 |
| 13 | 18 | 17 | 1 | 2 | 0 | 2 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 |
| 14 | 8 | 7 | 2 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 2 | 0 | 1 | 0 | 0 |
| 15 | 16 | 14 | 5 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |

Figure 4-10: Sample of extracted features using Message data parser (SMS)

## ➢ Email platform

1- The user was asked to login to his/her account on Microsoft Outlook, and the researcher installed an add-in software tool called *ReliefJet* in order to launch MS Outlook to export "sent item". This is available online at: https://www.reliefjet.com (as shown in Figure 4-11 below):

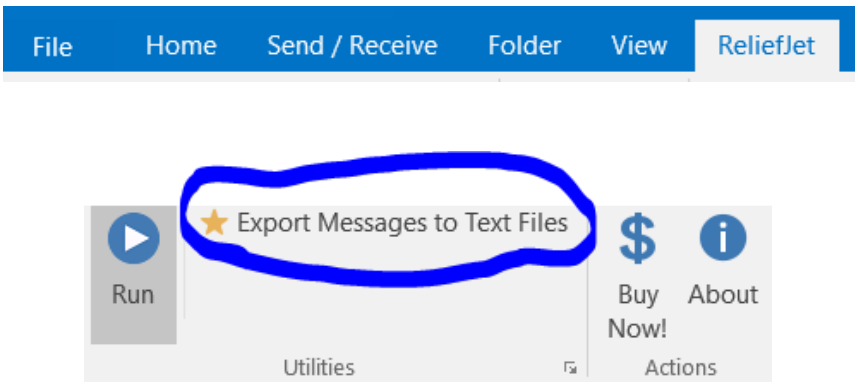


Figure 4-11: ReliefJet- MS Outlook add-ins Ribbon

2- The “sent items” option was selected, as shown in Figure 4-12 below. Then the sent Emails were saved in a file on the desktop as text files.

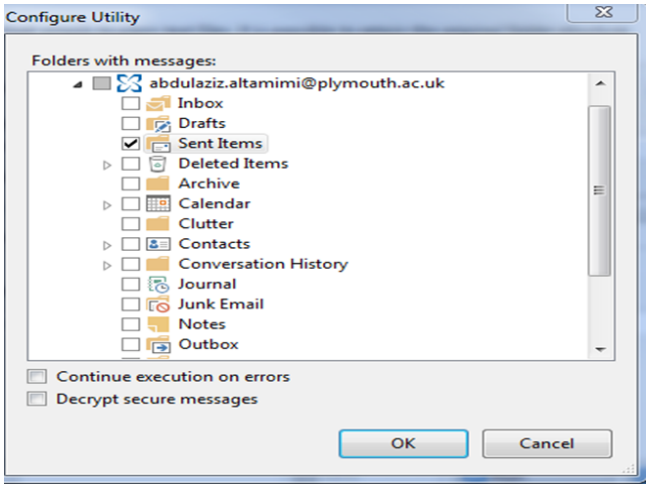



Figure 4-12: Selecting folder window

3- The data parser imported the text file to calculate the features from the text messages, which contained only numbers. The file of sent items has been deleted from the desktop and from the hard drive, and the user’s outlook Email was signed out and deleted before the user left; again, the only data saved was the calculated features. The results of the calculation features are shown in Figure 4-13 below:

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | F9 | F10 | F11 | F12 | F13 | F14 | F15 | F16 | F17 | F18 | F19 |
| 2 | 131 | 101 | 9 | 11 | 1 | 2 | 1 | 11 | 2 | 1 | 5 | 4 | 1 | 2 | 3 | 3 | 10 | 13 | 4 |
| 3 | 1060 | 971 | 58 | 79 | 17 | 42 | 27 | 130 | 20 | 10 | 43 | 67 | 1 | 4 | 43 | 34 | 59 | 56 | 26 |
| 4 | 356 | 306 | 38 | 30 | 1 | 20 | 10 | 29 | 8 | 6 | 10 | 22 | 0 | 3 | 11 | 10 | 24 | 20 | 8 |
| 5 | 406 | 358 | 39 | 32 | 2 | 25 | 17 | 30 | 8 | 4 | 10 | 28 | 1 | 3 | 13 | 12 | 27 | 28 | 7 |
| 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Figure 4-13 : Sample of extracted features using Message data parser (Email)

### ➢ Facebook platform

1- The user was asked to click on the Facebook icon on the main screen of the software data parser, and then a browser with a Facebook Graph Explorer

Tab would appear. On this tab, the user pressed “Log In” and entered their Facebook credentials, as shown in Figure 4-14 below.

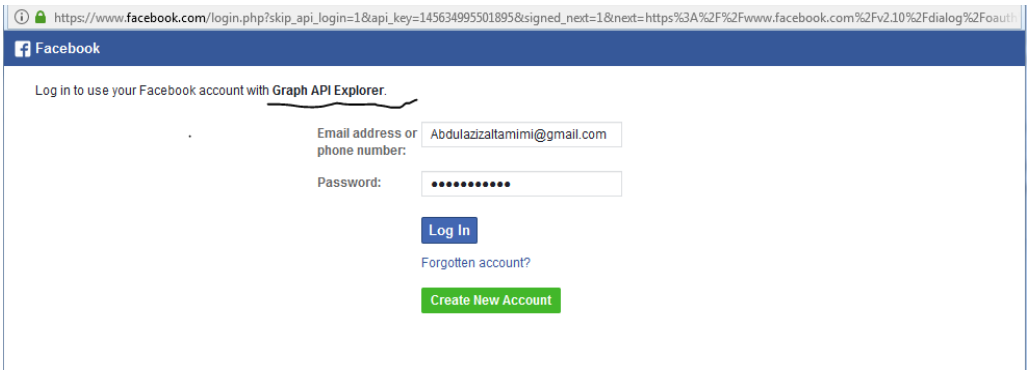



Figure 0-1: Facebook credentials

2- The user then pressed “Get token” -> “Get user token” and tick “*user_posts*”, before clicking the “Get access token” button, as shown in Figure 4 -15 below.

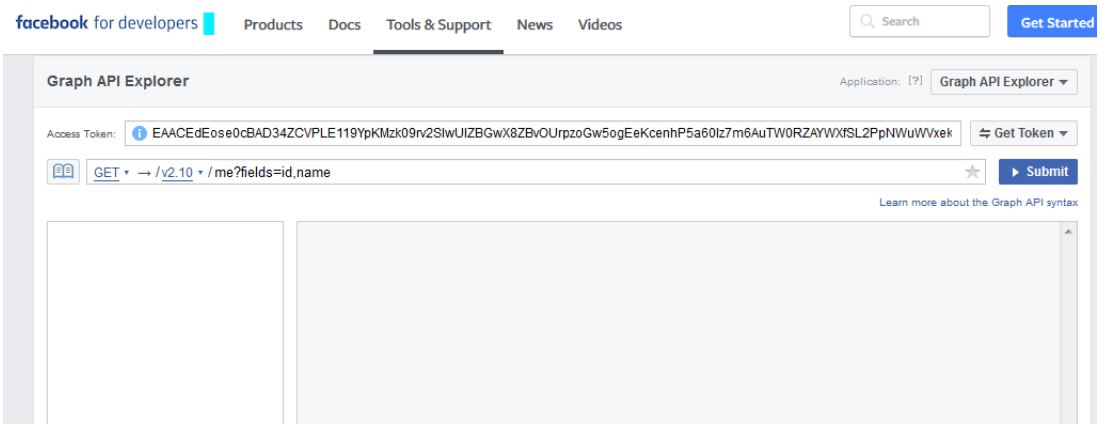



Figure 4-15: “Get access token” button

3- Next, the user ticked “*user_posts*”, and clicked the “*Get access token*” button, as shown in Figure 4-16 below.

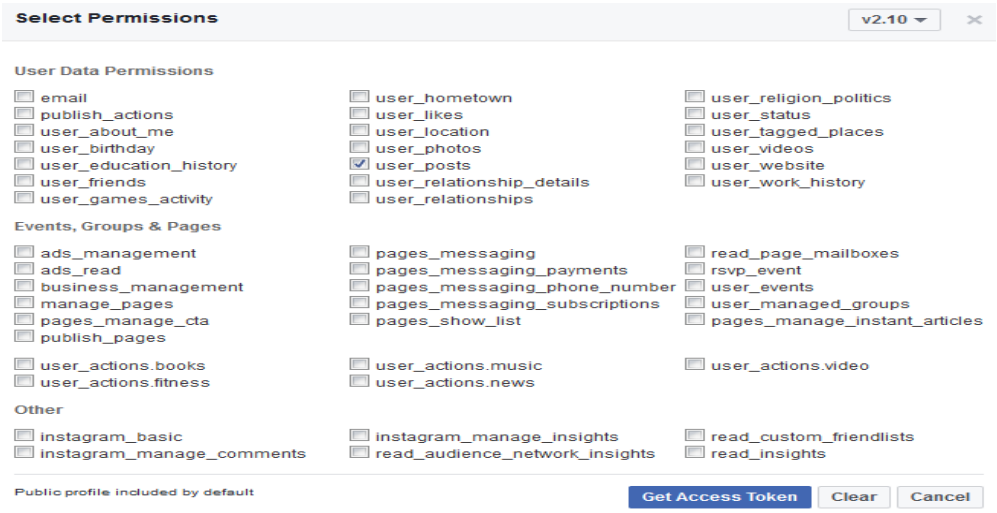



Figure 4-16: Selecting (user_ posts) option

4- The user token appeared in the "*Access Token*" text field, as shown in Figure 4-17.

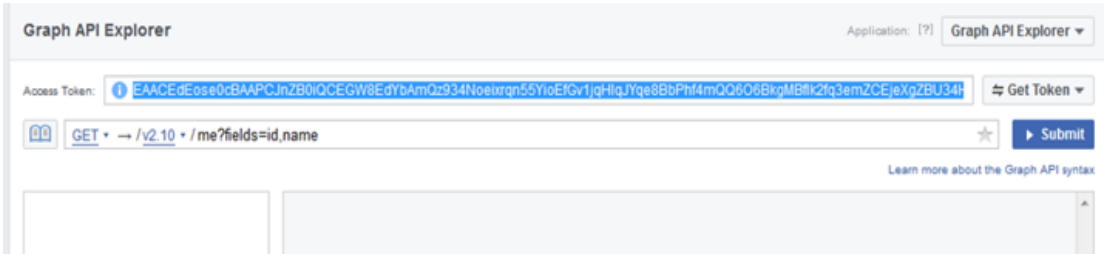



Figure 4-17: User token

5- The token appeared in the "Access Token", as shown in the above


6- The screenshot in Figure 4-18 shows that the researcher copied the data from there, closed the tab and pasted the token into the corresponding field of the window and pressed "Select".

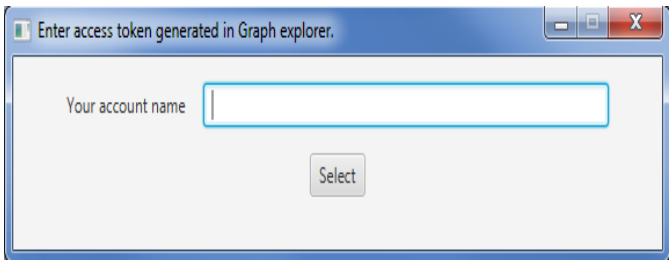



Figure 4-18: Corresponding field of the window

7- The user's Facebook account was signed out of before the user left, so that the only data obtained was the calculated features and the output file showing the calculated features, as shown in Figure 4-19 below.

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | F9 | F10 | F11 | F12 | F13 | F14 | F15 |
| 2 | 101 | 20 | 3 | 2 | 0 | 0 | 2 | 3 | 0 | 1 | 2 | 1 | 0 | 0 | 2 |
| 3 | 15 | 5 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 2 |
| 4 | 31 | 27 | 1 | 2 | 0 | 0 | 0 | 4 | 0 | 2 | 2 | 3 | 0 | 1 | 0 |
| 5 | 50 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 29 | 17 | 3 | 3 | 0 | 0 | 0 | 1 | 0 | 0 | 3 | 1 | 0 | 0 | 0 |
| 7 | 27 | 20 | 3 | 1 | 0 | 0 | 0 | 3 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| 8 | 128 | 106 | 7 | 13 | 0 | 3 | 4 | 9 | 7 | 0 | 10 | 11 | 0 | 1 | 2 |
| 9 | 8 | 8 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 0 | 0 | 0 |

Figure 4-19 : Sample of extracted features using Message data parser (Facebook)

### ➢ Twitter platform

1- The user was requested to click on the Twitter icon on the main screen of the software data parser, and then enter their account name without the "@" sign and press the "Select" button. Then, the Twitter API Explorer continued scraping the user's Twitter data, as shown in Figure 4-20.

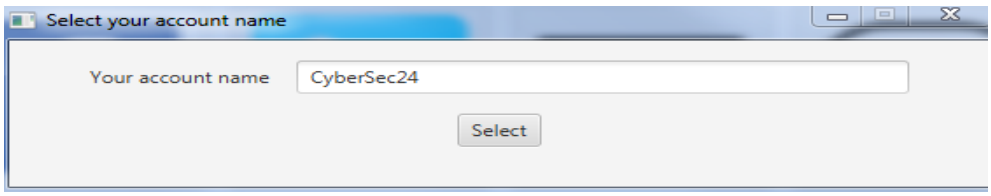



Figure 4-20 : User enters their account name without the "@" sign

2- The output file contained only the calculated features, and the user's Twitter account was signed out before they left; the only data obtained by the researcher was the calculated features, as shown in next Figure 4-21 below.

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | F9 | F10 | F11 | F12 | F13 | F14 | F15 |
| 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 19 | 5 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 2 |
| 4 | 20 | 14 | 2 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 2 | 1 | 0 | 0 | 2 |
| 5 | 8 | 5 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 2 |
| 6 | 120 | 102 | 7 | 9 | 1 | 3 | 6 | 9 | 3 | 0 | 7 | 11 | 1 | 0 | 5 |
| 7 | 99 | 72 | 5 | 3 | 0 | 1 | 0 | 10 | 1 | 0 | 6 | 5 | 0 | 1 | 2 |
| 8 | 5 | 5 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 2 |

Figure 4-21: Extracted features using Message data parser (Twitter)

# Appendix K –All users' EERs for Each Individually Platform

EER(%) _User-based experimental results _**SMS**_GB classifer_70/30 train/test (Top 100)

| User | EER | FAR | FRR | Threshold | Avg_EER | Avg_FAR | Avg_FRR |
|---|---|---|---|---|---|---|---|
| **1** | 5.580222 | 4.79393 | 6.366514 | 0.5 | 7.97 | 8.22 | 7.72 |
| **2** | 5.364103 | 5.6 | 5.128205 | 0.07 | | | |
| **3** | 0 | 0 | 0 | 0.01 | | 91.78 | 92.28 |
| **4** | 12.68829 | 12.97989 | 12.39669 | 0.28 | | acc. | 92.03 |
| **5** | 6.835938 | 7.03125 | 6.640625 | 0.15 | | | |
| **6** | 12.49002 | 12.32877 | 12.65127 | 0.37 | | | |
| **7** | 11.779 | 11.73533 | 11.82266 | 0.25 | | | |
| **8** | 12.49762 | 12.56281 | 12.43243 | 0.23 | | | |
| **9** | 16.2455 | 16.39344 | 16.09756 | 0.37 | | | |
| **10** | 12.73433 | 12.83906 | 12.62959 | 0.36 | | | |
| **11** | 4.908573 | 4.910058 | 4.907088 | 0.48 | | | |
| **12** | 6.410256 | 6.153846 | 6.666667 | 0.01 | | | |
| **13** | 3.571429 | 7.142857 | 0 | 0.01 | | | |
| **14** | 13.19497 | 13.39806 | 12.99188 | 0.43 | | | |
| **15** | 0 | 0 | 0 | 0.01 | | | |
| **16** | 5.411836 | 5.454545 | 5.369128 | 0.32 | | | |
| **17** | 14.16923 | 13.98157 | 14.3569 | 0.46 | | | |
| **18** | 7.326106 | 7.29927 | 7.352941 | 0.18 | | | |
| **20** | 4.524379 | 4.475588 | 4.573171 | 0.49 | | | |
| **24** | 13.6937 | 13.30798 | 14.07942 | 0.41 | | | |
| **25** | 8.81287 | 8.87574 | 8.75 | 0.09 | | | |
| **27** | 0 | 0 | 0 | 0.01 | | | |
| **30** | 3.913074 | 3.883495 | 3.942652 | 0.23 | | | |
| **38** | 3.721094 | 3.888889 | 3.553299 | 0.1 | | | |
| **39** | 14.15847 | 14.42155 | 13.89539 | 0.46 | | | |
| **40** | 7.211538 | 10.25641 | 4.166667 | 0.01 | | | |

EER(%) _User-based experimental results _**Twitter**_GB classifer_70/30 train/test (Top 275)

| User | EER | FAR | FRR | Threshold | Avg_EER | Avg_FAR | Avg_FRR |
|---|---|---|---|---|---|---|---|
| **1** | 17.82728 | 11.23596 | 24.4186 | 0.48 | 20.28 | 20.51 | 20.06 |
| **2** | 0 | 0 | 0 | 0.01 | | | |
| **3** | 21.38224 | 21.14286 | 21.62162 | 0.49 | | | |
| **4** | 32.47225 | 32.38636 | 32.55814 | 0.44 | | | |
| **5** | 33.32792 | 32.94798 | 33.70787 | 0.47 | | | |
| **6** | 30.41758 | 24.57143 | 36.26374 | 0.49 | | | |
| **7** | 31.35593 | 31.63842 | 31.07345 | 0.5 | | | |
| **8** | 17.8811 | 16.25 | 19.5122 | 0.5 | | | |
| **9** | 26.27119 | 25.9887 | 26.55367 | 0.47 | | | |

| | | | | |
|---|---|---|---|---|
| **10** | 23.86128 | 23.80952 | 23.91304 | 0.05 |
| **11** | 20.60606 | 20 | 21.21212 | 0.06 |
| **12** | 25 | 25 | 25 | 0.29 |
| **13** | 28.32081 | 25.56818 | 31.07345 | 0.5 |
| **14** | 25.97737 | 16.27907 | 35.67568 | 0.49 |
| **15** | 18.15925 | 18.28571 | 18.03279 | 0.48 |
| **16** | 28.15249 | 29.03226 | 27.27273 | 0.03 |
| **17** | 28.34186 | 22.34637 | 34.33735 | 0.5 |
| **18** | 0 | 0 | 0 | 0.01 |
| **19** | 25.33488 | 23.29545 | 27.3743 | 0.49 |
| **21** | 12.06395 | 11.62791 | 12.5 | 0.11 |
| **22** | 34.18835 | 33.51955 | 34.85714 | 0.49 |
| **23** | 13.20755 | 13.20755 | 13.20755 | 0.15 |
| **26** | 24.71509 | 24.57143 | 24.85876 | 0.5 |
| **28** | 34.18079 | 32.76836 | 35.59322 | 0.5 |
| **29** | 10.52632 | 10.52632 | 10.52632 | 0.05 |
| **30** | 0 | 0 | 0 | 0.01 |
| **31** | 17.33417 | 17.64706 | 17.02128 | 0.12 |
| **32** | 4.941176 | 4 | 5.882353 | 0.02 |
| **33** | 19.80392 | 20 | 19.60784 | 0.43 |
| **34** | 22.68174 | 22.47191 | 22.89157 | 0.47 |
| **35** | 15.31339 | 26.92308 | 3.703704 | 0.01 |
| **36** | 32.28051 | 32.58427 | 31.97674 | 0.46 |
| **37** | 34.59248 | 31.6092 | 37.57576 | 0.5 |
| **39** | 8.333333 | 8.333333 | 8.333333 | 0.09 |
| **41** | 31.99336 | 31.42857 | 32.55814 | 0.47 |
| **42** | 12.5 | 25 | 0 | 0.01 |
| **43** | 18.12189 | 17.91045 | 18.33333 | 0.46 |
| **44** | 12.5 | 25 | 0 | 0.01 |
| **45** | 7.142857 | 14.28571 | 0 | 0.01 |
| **48** | 23.36704 | 23.25581 | 23.47826 | 0.46 |
| **49** | 7.142857 | 14.28571 | 0 | 0.01 |

EER(%) _User-based experimental results _**Facebook**_GB classifer_70/30 train/test (Top 20)

| User | EER | FAR | FRR | Threshold | Avg_EER | Avg_FAR | Avg_FRR |
|---|---|---|---|---|---|---|---|
| **1** | 20.86957 | 21.73913 | 20 | 0.44 | 23.78 | 21.97 | 25.6 |
| **2** | 17.70833 | 18.75 | 16.66667 | 0.07 | | | |
| **3** | 11.80556 | 11.11111 | 12.5 | 0.01 | | 78.03 | 74.4 |
| **4** | 6.25 | 12.5 | 0 | 0.01 | | acc. | 76.215 |

| | | | | |
|---|---|---|---|---|
| **5** | 27.26262 | 26.86567 | 27.65957 | 0.34 |
| **6** | 38.87363 | 39.28571 | 38.46154 | 0.3 |
| **7** | 20.71429 | 20 | 21.42857 | 0.27 |
| **8** | 20.90571 | 19.23077 | 22.58065 | 0.1 |
| **9** | 26.12086 | 25.92593 | 26.31579 | 0.06 |
| **10** | 39.00966 | 39.13043 | 38.88889 | 0.08 |
| **11** | 14.93056 | 11.11111 | 18.75 | 0.5 |
| **12** | 17.39505 | 13.51351 | 21.2766 | 0.48 |
| **13** | 21.63743 | 22.22222 | 21.05263 | 0.47 |
| **14** | 36.08903 | 23.52941 | 48.64865 | 0.47 |
| **15** | 30.17647 | 28 | 32.35294 | 0.44 |
| **16** | 32.7957 | 32.25806 | 33.33333 | 0.37 |
| **17** | 4.805492 | 4.347826 | 5.263158 | 0.26 |
| **18** | 16.52174 | 13.04348 | 20 | 0.48 |
| **19** | 31.5873 | 14.28571 | 48.88889 | 0.43 |
| **20** | 27.61438 | 27.77778 | 27.45098 | 0.4 |
| **21** | 16.71959 | 16.41791 | 17.02128 | 0.46 |
| **22** | 26.13636 | 27.27273 | 25 | 0.1 |
| **23** | 22.87263 | 14.63415 | 31.11111 | 0.48 |
| **24** | 12.5 | 12.5 | 12.5 | 0.03 |
| **25** | 19.23624 | 19.11765 | 19.35484 | 0.45 |
| **26** | 21.8254 | 21.42857 | 22.22222 | 0.31 |
| **27** | 27.62987 | 25.71429 | 29.54545 | 0.46 |
| **28** | 19.44444 | 22.22222 | 16.66667 | 0.03 |
| **29** | 6.458333 | 6.25 | 6.666667 | 0.16 |
| **31** | 24.19908 | 15.78947 | 32.6087 | 0.47 |

| | | | | |
|---|---|---|---|---|
| **32** | 25.6305 | 25.80645 | 25.45455 | 0.35 |
| **33** | 33.33333 | 33.33333 | 33.33333 | 0.01 |
| **34** | 5 | 0 | 10 | 0.01 |
| **35** | 30.49563 | 28.78788 | 32.20339 | 0.48 |
| **36** | 37.5 | 33.33333 | 41.66667 | 0.36 |
| **37** | 33.25397 | 30.95238 | 35.55556 | 0.5 |
| **38** | 8.391608 | 9.090909 | 7.692308 | 0.49 |
| **39** | 6.25 | 0 | 12.5 | 0.04 |
| **40** | 36.81481 | 29.62963 | 44 | 0.3 |
| **41** | 35.64706 | 36 | 35.29412 | 0.34 |
| **42** | 33.40108 | 27.77778 | 39.02439 | 0.5 |
| **43** | 24.15949 | 24.59016 | 23.72881 | 0.31 |
| **46** | 36.38889 | 36.11111 | 36.66667 | 0.19 |
| **47** | 29.50311 | 30.43478 | 28.57143 | 0.04 |
| **48** | 30.49536 | 31.57895 | 29.41176 | 0.26 |
| **50** | 27.61608 | 27.02703 | 28.20513 | 0.36 |

EER(%) _User-based experimental results _**Email**_GB classifer_70/30 train/test (Top 20)

| User | EER | FAR | FRR | Threshold | Avg_EER | Avg_FAR | Avg_FRR |
|---|---|---|---|---|---|---|---|
| **1** | 10 | 8 | 12 | 0.43 | 12.03 | 11.45 | 12.61 |
| **2** | 12.40409 | 11.76471 | 13.04348 | 0.37 | | | |
| **3** | 25.15528 | 21.73913 | 28.57143 | 0.47 | | 88.55 | 87.39 |
| **4** | 11.25 | 12.5 | 10 | 0.49 | | Acc. | 87.97 |
| **5** | 13.37421 | 11.96581 | 14.78261 | 0.5 | | | |
| **6** | 25 | 0 | 50 | 0.01 | | | |
| **7** | 17.2619 | 16.66667 | 17.85714 | 0.22 | | | |
| **8** | 16.66667 | 13.33333 | 20 | 0.4 | | | |
| **9** | 16.69565 | 16 | 17.3913 | 0.04 | | | |
| **10** | 12.91667 | 12.5 | 13.33333 | 0.19 | | | |
| **11** | 4.545455 | 9.090909 | 0 | 0.01 | | | |
| **12** | 7.409274 | 7.526882 | 7.291667 | 0.35 | | | |

| | | | | |
|---|---|---|---|---|
| **13** | 15 | 16.66667 | 13.33333 | 0.08 |
| **14** | 3.846154 | 0 | 7.692308 | 0.49 |
| **15** | 25.32051 | 25 | 25.64103 | 0.1 |
| **16** | 7.738095 | 8.333333 | 7.142857 | 0.11 |
| **17** | 13.80952 | 14.28571 | 13.33333 | 0.46 |
| **18** | 9.722222 | 11.11111 | 8.333333 | 0.01 |
| **19** | 2.986907 | 3.846154 | 2.12766 | 0.01 |
| **20** | 8.391608 | 7.692308 | 9.090909 | 0.2 |
| **21** | 0 | 0 | 0 | 0.32 |
| **22** | 0 | 0 | 0 | 0.37 |
| **23** | 17.14286 | 14.28571 | 20 | 0.01 |
| **24** | 4.545455 | 9.090909 | 0 | 0.01 |
| **25** | 0 | 0 | 0 | 0.01 |
| **26** | 20.83333 | 16.66667 | 25 | 0.15 |
| **27** | 8.333333 | 8.333333 | 8.333333 | 0.49 |
| **28** | 15.89474 | 15.78947 | 16 | 0.4 |
| **29** | 16.66667 | 17.94872 | 15.38462 | 0.03 |
| **30** | 8 | 8 | 8 | 0.04 |
| **31** | 5.555556 | 11.11111 | 0 | 0.01 |
| **32** | 5.811688 | 5.714286 | 5.909091 | 0.42 |
| **33** | 11.80556 | 12.5 | 11.11111 | 0.3 |
| **34** | 15 | 15 | 15 | 0.01 |
| **35** | 3.794432 | 2.729529 | 4.859335 | 0.5 |
| **36** | 15.90909 | 16.66667 | 15.15152 | 0.45 |
| **37** | 11.80556 | 12.5 | 11.11111 | 0.01 |
| **40** | 25.54348 | 26.08696 | 25 | 0.06 |
| **41** | 16.26984 | 16.66667 | 15.87302 | 0.47 |
| **42** | 13.80952 | 13.33333 | 14.28571 | 0.21 |
| **43** | 17.06349 | 5.555556 | 28.57143 | 0.38 |
| **44** | 6.25 | 12.5 | 0 | 0.01 |
| **45** | 8.333333 | 16.66667 | 0 | 0.01 |
| **46** | 14.29739 | 14.70588 | 13.88889 | 0.24 |
| **47** | 20.88123 | 19.54023 | 22.22222 | 0.5 |
| **49** | 16 | 12 | 20 | 0.43 |
| **50** | 6.46357 | 6.741573 | 6.185567 | 0.21 |