Detecting Deceptive Behaviour in the Wild:

Text Mining for Online Child Protection in the Presence of Noisy and Adversarial Social Media Communications

Claudia Peersman, MLingLit. (Hons)

School of Computing and Communications

This thesis is submitted for the degree of
Doctor of Philosophy

MAY 2018

DECLARATION

I Hereby declare that, except where specific reference is made to the work of others, the contents of this thesis are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This thesis is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This thesis contains fewer than 80,000 words.

Claudia Peersman, MLingLit. (Hons)
May 2018

Abstract

This thesis, titled "*Detecting Deceptive Behaviour "in the wild": Text Mining for Online Child Protection in the Presence of Noisy and Adversarial Social Media Communications*", was submitted by Claudia Peersman (MLingLit., Hons) for the degree of Doctor of Philosophy in May 2018.

A real-life application of text mining research "in the wild", i.e. in online social media, differs from more general applications in that its defining characteristics are both domain and process dependent. This gives rise to a number of challenges of which contemporary research has only scratched the surface. More specifically, a text mining approach applied in the wild typically has no control over the dataset size. Hence, the system has to be robust towards limited data availability, a variable number of samples across users and a highly skewed dataset. Additionally, the quality of the data cannot be guaranteed. As a result, the approach needs to be tolerant to a certain degree of linguistic noise. Finally, it has to be robust towards deceptive behaviour or adversaries.

This thesis examines the viability of a text mining approach for supporting cybercrime investigations pertaining to online child protection. The main contributions of this dissertation are as follows. A systematic study of different aspects of methodological design of a state-of-the-art text mining approach is presented to assess its scalability towards a large, imbalanced and linguistically noisy social media dataset. In this framework, three key automatic text categorisation tasks are examined, namely the feasibility to (*i*) identify a social network user's age group and gender based on textual information found in only one single message; (*ii*) aggregate predictions on the message level to the user level without neglecting potential clues of deception and detect false user profiles on social networks and (*iii*) identify child sexual abuse media among thousands of legal other media, including adult pornography, based on their filename. Finally, a novel approach is presented that combines age group predictions with advanced text clustering techniques and unsupervised learning to identify online child sex offenders' grooming behaviour.

The methodology presented in this thesis was extensively discussed with law enforcement to assess its forensic readiness. Additionally, each component was evaluated on actual child sex offender data. Despite the challenging characteristics of these text types, the results show high degrees of accuracy for false profile detection, identifying grooming behaviour and child sexual abuse media identification.

ACKNOWLEDGEMENTS

This work was funded by Lancaster University through the European Commission Safer Internet Programme project (SI-2010-TP-2601002), *iCOP: Identifying and Catching Originators in Peer-to-Peer Networks*, and by the University of Antwerp (Belgium), *DAPHNE: Defending Against Paedophiles in Heterogeneous Network Environments*.

There are a number of people without whom this thesis might not have been written, and to whom I am greatly indebted.

First, I would like to express my endless gratitude to Professor Awais Rashid for his continuous support of my PhD study and related research, for his willingness to share skills, knowledge and expertise, for all his constructive feedback, for his guidance in setting and meeting both personal and professional goals and for conveying his never-ending enthusiasm in the field. I could not have imagined having a better mentor and supervisor.

Secondly, my supervisor, Dr. Alistair Baron, whom I am thankful for accepting the role of my supervisor after Prof. Rashid's transfer to Bristol University, and Prof. Paul Rayson and Prof. Brian Levine — my viva examiners — for their thoughtful comments and efforts towards improving this thesis.

Thirdly, my colleagues at the Cyber Security Group in Bristol, the DAPM and the CyBOK project and my former colleagues at Security Lancaster, the iCOP project and the University of Antwerp. I have learnt a great deal through your work.

Also, the people I have met from Belgian law enforcement, Interpol and the UNODC who have dedicated their professional lives to safeguarding children online. Without their support and guidance this study could never have been established. Your work, persistence and enthusiasm are truly inspiring.

I would like to dedicate this work to my parents, who have been a source of encouragement and inspiration to me throughout my life. Without their love and support, I could never have achieved this. Also, a very special thank you for all the extra help you provided during these last months of writing.

To my husband and friend for many years, Tom, thank you for your professional and energetic support when helping me collecting data for this thesis; a task which I could not have achieved alone. Furthermore, your practical and emotional support at home played a key role in the more difficult and demanding phases of this work.

Finally, saving the best for last, to Robbe, Niel and Bo. You have made me stronger, better and more fulfilled than I could have ever imagined. You boys mean the world to me. I truly hope my work can contribute to a safer Internet for you to enjoy.

TABLE OF CONTENTS

List of Tables

List of Figures

CHAPTER 1

INTRODUCTION

All words have the "taste" of a profession, a genre, a tendency, a party, a particular work, a particular person, a generation, an age group, the day and hour. Each word tastes of the context and contexts in which it has lived its socially charged life [11].

In recent years, several new media have seen the light of day, enabling millions of users to develop and support their personal and professional connections. Future generations will probably not be able to imagine a world without sharing their thoughts, experiences, pictures and videos with other users across the globe through applications such as social networking sites, peer-to-peer networks, virtual worlds and online gaming. In 2018, more than 4 billion people around the world have access to the Internet and over 3 billion of them are active on social media [104]. However, the ever increasing popularity of such environments inevitably attracts criminals, who exploit social media to approach potential victims.

A recent survey organised by the EU Kids Online project showed that children are especially vulnerable to the risks of using the Internet: 9- to 16-year-olds spend 88 minutes a day online on average, with 49% of these adolescents going online in their bedroom — away from adult supervision. Moreover, the study found that 34% of the children who took part in the survey had added people to their social media friends lists they had never met face-to-face, 15% had sent

personal information, pictures or videos of themselves to strangers and 9% had agreed to meet in person with someone they had only met online [130].

Based on a study by the European Parliament [96], four key risk factors can be identified that specifically relate to children and the Internet: (*i*) a child can be exposed to harmful and illegal content; (*ii*) social media platforms can be utilised to solicit children, which is often the prelude for child sexual exploitation; (*iii*) children can become engaged in illegal activities, such as disseminating child abuse media or coercing other children into becoming victims; and (*iv*) the over-abundance of virtual communities that enable offenders to remain anonymous or to create false virtual identities facilitates the production and dissemination of child sexual abuse media (CSAM).

To combat the current and future challenges of such types of cybercrime, the authors of [94] already argued that there is a need to re-examine standard digital forensic procedures and the use of investigative technology by incorporating *intelligent* technologies, i.e. techniques from artificial intelligence, computational modelling and social network analysis. In this thesis, we investigate if such intelligent techniques can be used to focus police investigations pertaining to child protection and reduce the amount of time spent looking for digital evidence. More specifically, this study combines the advantages of text mining and machine learning procedures (both AI) with insights from social science, such as (socio)linguistic theory, criminology and psychology, to automatically detect deceptive behaviour "in the wild", i.e. in online social media (see [115]).

Such a real-life application of text mining research differs from more general applications in that its defining characteristics are both domain and process dependent. This difference gives rise to a number of challenges of which contemporary research has only scratched the surface. More specifically, a text mining approach applied on social media communications typically has no control over the dataset size. Hence, the system has to be robust towards limited data availability, a variable number of samples across users and a highly skewed dataset. Additionally, the quality of the data cannot be guaranteed. As a result, the approach needs to be tolerant to a certain degree of linguistic noise. Finally, it has to be robust towards deceptive or *adversarial* behaviour, i.e. users who attempt to hide their criminal intentions (*obfuscation*) or who assume a false digital persona (*imitation*) [32].

In this dissertation, we focus on crucial aspects of methodological design that significantly affect the performance of a text mining approach when applied “in the wild", which is still a lacuna in the field. Additionally, we present a novel approach for automatically identifying adversarial behaviour under the complex conditions described above. Given the devastating, lifelong impact, both physically and emotionally, of child sexual abuse — both with regard to the victims and their family — and the lack of technologies designed to help detect online child sex offenders [15, 57, 93, 162], we investigate these objectives within the framework of designing text-based, investigative approaches for child protection purposes. In particular, the key contributions of this thesis are as follows.

- **A collection of new corpora**. This study is based on three different Flemish Dutch datasets[1] containing social media communications that were collected in the context of the DAPHNE project[2] and one multilingual dataset of filenames shared on P2P networks that was gathered in the framework of the iCOP project[3]. More specifically, the first corpus contains a total of 1.4 million Flemish Dutch postings that were provided by the Belgian social networking site Netlog[4] between 2010 and 2014. Each message contains meta-information on the user's profile, such as age, gender and location. Because Netlog was very popular among adolescents at the time of data collection, over 80% of the messages in the NETLOG corpus were produced by teenagers between eleven and twenty years old. Given the scarcity of such available data [15], this corpus is a valuable resource for both computational and sociolinguistic research. Secondly, we set up an adversarial stylometry study[5] for which we recruited forty-eight volunteers (the one half adults and the other adolescents) to participate in a private one-on-one chat room conversation, i.e. the VOLUNTEER corpus. For the purposes of this thesis, the adults were asked to pose as adolescents. Thirdly, we manually collected a dataset of actual offender-victim online conversations from recently closed court case files in collaboration with Belgian law enforcement, i.e. the PREDATOR

[1] The author of this thesis is a native speaker of Flemish Dutch.

[2] The author conducted research at the University of Antwerp (Belgium) and Lancaster University (UK) during this project.

[3] The iCOP project was led by Lancaster University (UK).

[4] `http://nl.netlog.com/`

[5] We conducted this study in accordance with the ethical standards and ethics guidelines specified by the University of Antwerp (see Appendix A).

corpus. Finally, the fourth dataset consisted of non-CSAM filenames obtained from various web sources like *flickr.com* and *youtube.com* and actual child sexual abuse media filenames, which we collected from the case files mentioned above and which were provided to us by Interpol[6].

- **Automatic age and gender prediction in online social networks**. Compared to prior research on literary works, dealing with social media communications typically implies analysing a large variety of short text samples. This focus is more challenging, because each dataset contains a great amount of different linguistic features, but only few of these features are present in each sample. Additionally, linguistic noise, such as abbreviations, non-standard language use and spelling variations and errors, is innate to social media communications and poses great difficulties for off-the-shelf Natural Language Processing (NLP) applications. In this dissertation, we provide a systematic study of different aspects of methodological design incorporated in a state-of-the-art user profiling approach to assess its viability when applied "in the wild".

- **Identifying false user profiles**. Previous work [177] already stated that creating a false digital persona is one of the key tactics employed by cyber criminals to establish contact with potential victims. Examples of such criminal exploitation of false user profiles include child sex offenders pretending to be adolescents to gain their victims' trust [74]; mass-marketing fraud, such as advance fee scams and online dating scams, in which (multiple) fake identities are created and misused for financial gain [58, 223]; and online recruitment of people to fight for radical causes [218]. In this thesis, we evaluate the performance of a novel user profiling system when confronted with *imitation*, i.e. social network messages produced by adults imitating a younger age group.

- **Detecting criminal media.** Recent work on identifying criminal media shared on P2P networks showed that offenders tend to include linguistic noise in their filenames as an adversarial tactic to *obfuscate* the illegal content of their files and circumvent detection

[6]We collected all filenames in accordance with the ethical standards and ethics guidelines specified by Lancaster University.

by law enforcement, while maintaining their availability to other offenders [124]. This dissertation presents an intelligent solution to this challenge, which adopts text mining techniques to determine the likelihood that a candidate file contains criminal content based on its filename.

- **Grooming detection.** There is a consensus among experts that a grooming process, in which a child sex offender attempts to seduce his or her victims and prepare them for the actual abuse, consists of multiple, recurrent stages. In this thesis, we explore the use of unsupervised learning to develop a new methodology to automatically identify different stages of grooming.

This introductory chapter starts with describing the context of this dissertation (Section 1.1). Following, we introduce the research objectives in Section 1.2 and summarise the main contributions of this thesis in Section 1.3. A chapter guide is provided in Section 1.4. Finally, we present an overview of the publications that have emerged from this thesis in Section 1.5.

1.1 Context

This thesis presents a novel approach to assist law enforcement in their investigations pertaining to online child sexual abuse, while also performing basic research in the fields of text mining and computational stylometry. This section introduces the intelligent technologies used in this dissertation. An overview of the current trends and forms of online child sex exploitation can be found in Chapter 6, Section 6.2.1.

1.1.1 Intelligent Technologies

In this section, we introduce the specific areas from the fields that are most relevant to the research presented in this dissertation. A more detailed survey of the methodology used in this study can be found in Chapter 2.

1.1.1.1 Text Mining

With the increasing availability of large amounts of digital text, detecting patterns and trends manually has become increasingly challenging. A popular technique to resolve these issues is text mining. A text mining approach is typically designed within an NLP framework, as a level of information extraction from text. Its main objective is to build an intelligent tool, that has the capability of analysing large amounts of natural language texts (e.g., newspaper articles, books or emails) and extracting useful information in a timely manner. Hence, it is a step forward from the information retrieval task, in which the best matches in a database are calculated based on a user query, to a level of exploring the various types of high quality knowledge that can be extracted from text. Although this is a relatively new research area, the technology is already being used in a wide variety of applications, such as biomedical applications (e.g., GoPubMed[7], a knowledge-based search engine for biomedical texts), business and marketing applications (e.g., stock return prediction [71]), security applications (e.g., automatic monitoring of Internet news, blogs and social media [230]) and academic applications (e.g., academic publishers making their papers available for text mining purposes).

A text mining approach typically involves the following six steps:

1. A dataset of text documents relevant to the task at hand is collected.

2. Each document is pre-processed. More specifically, the data is converted to the desired format, is split up into individual words and punctuation marks (i.e., *tokenised*) and processed for removing content undesirable for the task in question (e.g., hyperlinks, non-standard word forms, redundancies and stop words). Most text mining approaches also include *stemming* or *lemmatisation* in their pre-processing procedure: stemming implies that each word in the dataset is reduced to its word stem, base or root form (for example, "studying" is reduced to "study" and "studies" or "studied" to "studi"), while lemmatisation refers to reducing each word to its lemma or dictionary form (for example, "studying", "studies" and "studied" are all represented by "study").

[7] `http://www.gopubmed.com/web/gopubmed/`

3. The documents are transformed from the full text version to a vector space model that represents the different sets of linguistic features present in each document (e.g., words, characters, Parts-of-Speech and semantic roles).

4. Statistical techniques are applied to determine which features are most informative for the task at hand. Usually, non-discriminative features are discarded by the system to reduce the dimensionality of the dataset.

5. The resulting structured database is analysed using either automatic classification or clustering techniques that are also used in data mining. In most cases, analysis happens using machine learning or statistical algorithms.

6. The output of the previous step is evaluated and can be stored or used in a series of following text mining experiments.

Figure 1.1 shows a standard text mining process flow. The next section introduces the emerging research field of computational stylometry, in which the relation between natural language and its users is typically studied by adopting such a text mining framework.

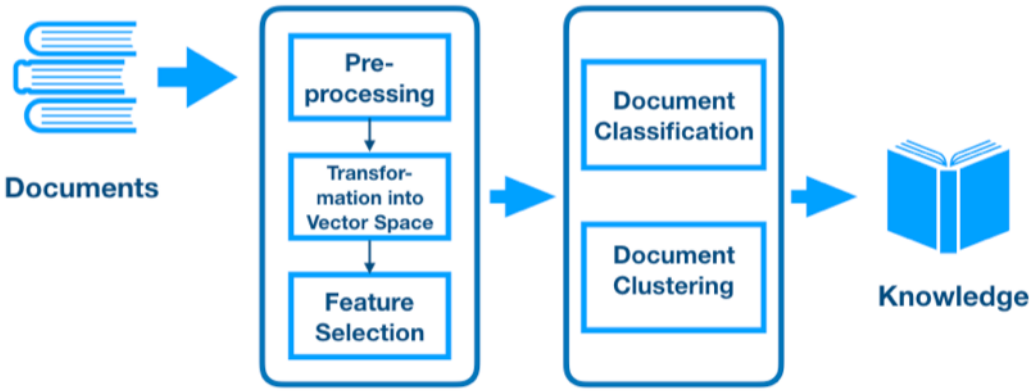

Figure 1.1: A standard text mining process flow.

1.1.1.2 Computational Stylometry

Language is a social phenomenon and language variation is, as a consequence, innate to its social nature. By selecting between a range of potential variations of language use, people construct a social identity, which can be employed to achieve certain social goals. In other words, language users can make use of specific language varieties to represent themselves in a certain way. This freedom of choice, which is shaped by a series of both consciously and unconsciously made linguistic decisions, is often referred to as *speaker agency*. Such variation can be manifested at various levels of language use, for example, the choice between different words, phonological variants or grammatical alternatives, and is typically influenced by a speaker's (intended) audience, demographic variables (such as age group, gender or background) and objectives (e.g., knowledge transfer, persuasion or likeability). While sociolinguistics mainly focuses on the reciprocal influence between language and society (see [38, 39, 55, 86, 119, 198]), stylometry studies are mainly based on the hypothesis that the combination of these (un)conscious linguistic decisions is unique for each individual — like a fingerprint or a DNA profile [50, 206] — and that language users (i.e., both speakers and authors) can be identified by analysing the specific properties of their linguistic choices. The idea of such a *human stylome* [206] can be dated as far back as the mediaeval scholastics[8].

In modern times, the approach of analysing a text on different linguistic levels to determine its authorship was adopted within research fields such as stylistics and literary criticism, one of the most prominent examples being the investigations into the literary works attributed to Shakespeare [54]. This type of research is commonly referred to as *traditional* authorship attribution and typically involves in-depth reading by human experts. However, in the late 19th century, a new line of research demonstrated that an author's writing style could be quantified. A study by [145], for example, showed that the authorship of the Federalist Papers could be settled by comparing the distribution of *stop words* (or *function words*) in the disputed texts to other texts written by the three candidate authors[9]. As a result, the twentieth century has witnessed a number of attempts to determining discriminative style markers in cases of disputed authorship

[8] During the Middle Ages, a text's veracity typically depended on the hand that produced its contents [36, 114].
[9] Alexander Hamilton, James Madison and John Jay.

(e.g., [229, 235]), even ranging to applications in criminal court cases [98].

The arrival of modern computational methods and the emergence of the Internet have instigated a new research area, combining insights from the fields of stylometry to techniques commonly used in computer science. Based on the assumption that authors can be distinguished by their stylome, *non-traditional* authorship attribution typically focuses on developing a computational model that can automatically identify the author of a given text. The dominant approach in these studies is typically based on text mining methods, which are used to automatically attribute one or more predefined thematic categories — such as authors— to a set of natural language texts (e.g., books, papers or emails)[10]. Recently, a significant part of the field has shifted focus from attributing texts to specific authors to investigating whether certain aspects in an author's writing style can be generalised for larger author groups belonging to, for example, the same age or gender group or showing similar personality traits (e.g., outgoing or withdrawn). Together with non-traditional authorship attribution, such *author profiling* studies constitute the rapidly developing field of computational stylometry [50].

In recent years, computational stylometry has evolved into a state where it is feasible to develop useful applications for large, formal text corpora. However, when analysing textual data produced on social media, researchers are confronted with several issues. First, such online communications typically consist of very short messages. Most computational stylometry studies have tackled this problem by including multiple messages per user in their experiments, resulting in text fragments ranging from 250 to several thousands of words on average per user. Additionally, previous work by, for example [1, 98, 99] has demonstrated that even high-performance authorship identification systems can be reduced to random behaviour when they are confronted with adversarial texts, i.e., passages that include obfuscation or imitation.

Contrary to prior work in the field, this thesis investigates the feasibility to develop a text-based user profiling system by incorporating only a single message per user. Furthermore, a systematic evaluation of the different aspects of experimental design is presented to provide insight into potential scalability issues of applying a text mining approach "in the wild". Finally, the system is stress-tested by applying it to adversarial text samples and actual offender-victim

[10]The history and background of authorship attribution studies can be found in [98].

communications. The next section describes the challenges that are dealt with in this work and states its key research objectives.

1.2 Research Objectives

This study differs from most contemporary work in the fact that it examines three aspects that are becoming increasingly important in the field and that are essential for designing a real-life application that can be used to support digital forensic investigations, namely: (*i*) dealing with linguistically noisy texts, (*ii*) sparse data analysis and (*iii*) detecting adversarial text samples. This section elaborates on each of these aspects and formulates the basic research questions for this study.

1.2.1 Noisy Data

The increased level of immediacy in computer-mediated communication has led to the rise of a new *glocal* language variety, displaying both characteristics from a global *Internet language* — also called *chatspeak* — leading to a wild proliferation of new language varieties (e.g., Internet abbreviations, acronyms, character flooding, concatenations and emoticons) and characteristics that are innate in people's local dialect [7, 49]. The presence of both these types of linguistic noise are said to provide a significant challenge for text mining research, because many off-the-shelf NLP tools (used for e.g., automatic stemming, lemmatisation or part-of-speech tagging) fail to correctly analyse this anomalous input. The standard practice in a text mining approach is to discard all non-standard language varieties (e.g., [154, 186, 226]) or to normalise them to improve feature extraction procedures (e.g., [19, 156]). However, systematic studies of the level of linguistic noise in (Flemish) online social network communications are still lacking in the field. Additionally, previous work in spoken discourse studies have observed strong correlations between the use of non-standard language and sociological variables such as age and gender (e.g., [39, 42, 55, 86, 119–122, 143, 198, 202, 216, 225]). Therefore, in this thesis, we investigate the potential of such (un)consciously made linguistic choices to contribute to a better performance of a user profiling system. More specifically, we examine the following research questions:

Q1 How linguistically noisy are Flemish social media communications?

Q2 What is the effect of age and gender on both types of non-standard language use (i.e. chatspeak and dialectal forms) in social media communications?

Q3 Can including such linguistic noise contribute to a higher performance of a text mining approach designed to perform user profiling in online social media?

1.2.2 Data Size

Standard practice in a wider text mining context that involves short text samples is to increase the data in each sample by, for example, grouping multiple text fragments written by the same author (e.g., [152, 186, 232]) or by incorporating additional word level concept information obtained from external sources, such as pre-trained word embeddings, WordNet, concept annotations or snippets produced by public search engines (e.g., [80, 129, 139]). However, for a real-life application, it is essential that a text mining approach is able to achieve a reliable performance, even when confronted with limited data availability. Furthermore, in a digital forensics context, it would be inconceivable to combine evidence with external content that was not produced by the person under investigation. Therefore, in this thesis we investigate the feasibility of designing a text mining approach that can be used with sufficient reliability on large datasets containing short text samples. The research questions addressed are as follows:

Q4 What is the effect of different feature selection, representation and machine learning techniques on the ability of a text mining approach to deal with highly sparse data?

Q5 Which feature types show more robustness under such complex conditions?

Q6 Which techniques can be used to increase the performance of the approach?

1.2.3 Adversarial Data

Contemporary computational stylometry research typically focuses on two aspects: (*i*) identifying and extracting linguistic features that are potentially discriminative for an author's *writing print* (or *stylome*) and (*ii*) developing an efficient computational model that includes these

features to automatically determine an author's identity or profile. Although a range of feature types and computational methods have been suggested for the task, to this date, the field is dominated by studies that evaluate their computational stylometry approaches on non-deceptive datasets. However, a key issue when designing a text mining approach to be used in cybercrime investigations is whether it will remain useful when it is confronted with adversarial behaviour. Therefore, in this study we evaluate the best performing models on (*i*) a dataset of adults posing as adolescents (*imitation*); (*ii*) a dataset containing chat room communications by recently convicted child sex offenders (*obfuscation/imitation*); and (*iii*) a dataset of actual CSAM filenames in which linguistic noise and specialised vocabulary are used as an adversarial tactic to circumvent detection by law enforcement (*obfuscation*). Within these adversarial experiments, the following research questions are examined:

Q7 Is it feasible to design a system that is able to identify a user's age group even if (s)he imitates the writing style of a different age group?

Q8 Which experimental design leads to a more robust performance when detecting adversarial text passages?

Q9 Can a text mining approach be used with sufficient reliability to identify media that contain child sexual abuse content based on linguistic features in its filenames, despite the even shorter text samples and obfuscation techniques used by offenders to avoid detection by law enforcement?

Q10 Can a text mining approach be used to identify different stages of grooming that are employed by child sex offenders to deceive their victims?

1.3 Contributions

In this thesis, we investigate the feasibility to perform automatic user profiling "in the wild", which is still a lacuna in the field. First, different aspects of experimental design that can potentially affect the efficiency of our approach when applied to short, noisy and adversarial data are placed under scrutiny. Additionally, we examine three automatic text categorisation

tasks, namely the feasibility to (*i*) identify a social network user's age group and gender based on textual information found in only one single message; (*ii*) aggregate predictions on the message level to the user level without neglecting potential clues of deception and detect false user profiles on social networks; and (*iii*) identify child sexual abuse media among thousands of legal other media, including adult pornography, based on their filename. Also, we present a novel approach that combines age group predictions with advanced text clustering techniques and unsupervised learning to identify grooming behaviour. Finally, systematic analyses of the effects of our approach when confronted with the challenges mentioned above allows us to evaluate if a state-of-the-art text mining approach is fit for application in a digital forensics context.

1.4 Chapter Guide

In the beginning of this chapter, we presented an outline of the context and motivation behind this work. Additionally, we introduced the key research questions, together with the text mining approach adopted in this study. The remainder of this chapter briefly describes the content of each of the following chapters of this dissertation and provides an overview of the publications emerging from this thesis.

Chapter 2 describes the methodology we adopt in this dissertation and provides a description of the approach in terms of the most important aspects of methodological design, such as feature types, feature selection algorithms, feature representation methods and machine learning techniques.

Chapter 3 introduces the main dataset we use in this study, i.e., the NETLOG corpus. Additionally, we analyse the level of linguistic noise attested in this dataset and we investigate potential correlations between a Netlog user's demographic features (age and gender) and the probability of producing non-standard language varieties.

Chapter 4 examines the feasibility to perform automatic user profiling in online social networks when confronted with short, linguistically noisy text samples. In this chapter, we also investigate whether including insights from previous sociolinguistic studies can increase the system's performance when detecting age group and gender.

In **Chapter 5**, we accumulate the best performing aspects of experimental design on the message level and aggregate them to the user level. The performance of this novel approach is compared to the more traditional user-based approach when applied on adversarial data.

Chapter 6 focuses specifically on identifying offender behaviour by examining the feasibility to automatically detect grooming and child sexual abuse media. We evaluate the approach on actual criminal data that were obtained in collaboration with law enforcement.

Finally, **Chapter 7** formulates an answer to the research questions that were introduced in this first chapter. Additionally, we summarise this thesis' key research contributions, discuss the implications for digital forensic applications of text mining in online social media and propose future research perspectives.

1.5 Publications Emerging from this Thesis

All work presented in this dissertation is that of the author, unless it is indicated otherwise. Some of the research presented in this thesis has been previously published, with the support of other authors, as described below.

1. The post-based approach described in Chapter 4 was previously published under the title 'Predicting age and gender in online social networks' [159]. Prof. Daelemans and Van Vaerenbergh provided guidance in evaluating the methodology.

   [159] C. PEERSMAN, W. DAELEMANS, AND L. VAN VAERENBERGH, *Predicting age and gender in online social networks*, in Proceedings of the 3rd international workshop on search and mining user-generated contents (ACM), 2011, pp. 37-44.

2. A less detailed description of the grooming detecting component was previously published under the title 'Conversation level constraints on pedophile detection in chat rooms' [163]. Together with the author, Dr. Vaassen and Dr. Van Asch designed the user-level experiments (which are not included in this thesis) and Prof. Daelemans provided insight in evaluating the methodology.

[163] C. PEERSMAN, F. VAASSEN, V. VAN ASCH, AND W. DAELEMANS, *Conversation level constraints on pedophile detection in chat rooms*, Notebook for PAN at CLEF, 2012.

3. The NETLOG corpus described in Chapter 3 and its characteristics have been previously introduced by the author in [107]. Dr. Kestemont was the instigator of the Chatty project, in which part of the dataset was coded to help develop NLP tools for Flemish chatspeak. Dr. De Decker, Dr. De Pauw, Dr. Luyckx, Dr. Morante, Dr. Vaassen, Dr. van de Loo and Prof. Daelemans contributed to the coding of the Netlog subset described in [107].

   [107] KESTEMONT, M., C. PEERSMAN, B. DE DECKER, G. DE PAUW, K. LUYCKX, R. MORANTE, F. VAASSEN, J. VAN DE LOO, AND W. DAELEMANS, *The Netlog Corpus: a resource for the study of Flemish Dutch Internet language*, in LREC, 2012, pp. 1569-1572.

4. The filename categorisation approach described in Chapter 6 was published in [161] and [162] by the author under the supervision of Prof. Rashid, who was the project leader of iCOP. Dr. Schulze designed the image analysis module, the interviews with law enforcement were carried out and analysed by Dr. Brennan and Dr. Fischer implemented the filename and image analysis modules into the final version of iCOP toolkit.

   [161] C. PEERSMAN, C. SCHULZE, A. RASHID, M. BRENNAN, AND C. FISCHER, *iCOP: Automatically identifying new child abuse media in p2p networks.*, in Proceedings of the Security and Privacy Workshops (IEEE), 2014, pp. 124-131.

   [162] C. PEERSMAN, C. SCHULZE, A. RASHID, M. BRENNAN, AND C. FISCHER, *iCOP: Live forensics to reveal previously unknown criminal media on P2P networks*, Digital Investigation, 18 (2016), pp. 50-61.

5. The background and related work on non-standard language variation described in Chapter 3 have been previously published in [160], together with the qualitative and quantitative analyses on a subset of the NETLOG corpus. Prof. Vandekerckhove, Dr. Vandekerckhove, Prof. Daelemans and Prof. Van Vaerenbergh provided insight and statistical expertise in evaluating the methodology.

[160] C. PEERSMAN, W. DAELEMANS, R. VANDEKERCKHOVE, B. VANDEKERCKHOVE, AND L. VAN VAERENBERGH, *The effects of Age, Gender and Region on non-standard linguistic variation in online social networks.*, arXiv preprint arXiv:1601.02431, 2016.

6. The iCOP toolkit and its filename categorisation component were analysed for its capability of supporting digital investigations by the author, under the supervision of Prof. Rashid, in [158]. Alongside the author, Dr. Pasquale, Dr. Tun, Dr. Alrajeh, Dr. Nuseibeh and Prof. Rashid defined the requirements for the forensic readiness of software systems described in chapter 7.

   [158] L. PASQUALE, T. TUN, D. ALRAJEH, B. NUSEIBEH, C. PEERSMAN, AND A. RASHID, *Towards forensic-ready software systems*, in Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results, 2018, pp. 9-12.

CHAPTER 2

BACKGROUND

2.1 Introduction

Text mining is an exciting research field that combines techniques from data mining, machine learning, natural language processing and information retrieval to provide solutions regarding the information overload in an increasingly digitalising world. This chapter describes the different methodological steps of the text mining approach this study is based on. More specifically, we focus on the different NLP techniques used to pre-process the datasets, together with the feature engineering, classification and clustering techniques presented in the following chapters. As a result, this overview is biased towards the techniques used in this dissertation. More general introductions to text mining methods can be found in [2, 64].

2.2 Feature Types

The first step in a text mining approach consists of pre-processing the textual data and extracting features that represent linguistic information on the character, lexical, syntactic or semantic level. This subsection introduces each of these types.

Lexical features are commonly used in text mining studies pertaining to social media user

profiling (e.g., [5, 13, 34, 66, 73, 116, 153, 176, 185, 186, 226, 232]). Not only are they relatively easy to extract from text — they only require tokenisation — they usually contain useful stylistic information. In contemporary research, two subtypes are distinguished: *content* words and *function* words. While content words carry the primary communicative message of an utterance (e.g., nouns, adjectives, verbs and adverbs), function words are used to express a grammatical or structural relationship between words. Examples of function words are determiners, ad-positions, pronouns, conjunctions, auxiliary verbs, interjections and particles. The advantage of the fact that the latter features bear little lexical meaning is that they are less under a user's conscious control and less topic-dependent (see [43, 148]). Once lexical features are extracted, the document is typically represented as a multi-set of its words. In text categorisation studies this is usually referred to as a *bag-of-words* (BOW) model. However, such a BOW model does not take into account any information on the context of the lexical features that are included in the model. As an alternative, some user profiling studies [5, 34, 176, 185] incorporate an n-gram model, i.e. a model that extracts a pre-defined n number of consecutive words, in their system.

Character features, from their part, represent a text by a pre-defined number of consecutive characters — character n-grams — enabling to capture information on other linguistic levels, such as pre- and suffix information, which could also be useful for tracing stylometric evidence. Despite their success in closely related fields such as automatic language identification (e.g., [37, 56]) and authorship attribution [40, 68, 84], character n-grams have only been applied to the user/author profiling task since the 2010s [34, 185].

Syntactic features, a third common feature type, are typically based on frequencies of a variety of syntactic constructions, that are extracted by natural language processing techniques. Prior work has used the output of text chunkers or parsers to improve the results for authorship attribution [40, 84, 205]. Some recent work has also suggested combining Part-of-Speech (POS) tags, such as nouns, adjectives, conjunctions, auxiliary verbs and prepositions, with other feature types [113, 206, 233]. The main advantage of such syntactic features is that — like function words — they are less topic-dependent than lexical or character n-gram features. However, because most off-the-shelf NLP tools are trained on standard language use, only few online user profiling studies included syntactic features in their work [186, 232].

Semantic features represent the basic conceptual components of meaning for any lexical item [70]. In other words, semantic features are used to explain how words are semantically related and how they differ in meaning. For example, words like "father", "mother", "brother" and "sister" all belong to the *family* category, but they belong to a different category in terms of generation or adulthood. To extract such features, the authors of [166] developed the Linguistic Inquiry and Word Count (LIWC) word category lexicon, which includes sixty-four different categories of language, among which topical categories ("family", "affect", "occupation", "body", etc.)[1]. Hence, the number of references to each category can be included as a feature. Because the lexicon only includes standard language forms, so far, only [66] has used them to predict gender in Twitter data.

A few other *non-traditional* features have been used for age or gender classification in computer-mediated communications, such as the use of non-dictionary words [73, 176, 186], hyperlinks [186], background colour, word fonts and cases, punctuation marks and emoticons [176, 226], average word and sentence length [73, 232], hashtags, retweeting frequency and followers/friends ratio [5].

This thesis investigates the usefulness of both traditional and non-traditional feature types in Chapters 4 to 6.

2.3 Feature Selection

When working with large datasets, analysing lexical features often results in a vast amount of potentially discriminative features. Because training on such a high dimensional dataset usually entails a computationally expensive training phase, a wide variety of feature selection methods have been developed that not only reduce the computational cost, but also — in some cases — result in more accurately trained models. Generally, there are two common approaches to performing feature selection: the *filter* and the *wrapper* approach. In the filter approach, features are first ranked based on a single feature relevance criterion for which no machine learning experiments are required. The simplest criterion to reduce the number of features in a dataset is to include the most frequent features in the training data. However, words that are frequently

[1]This lexicon was also developed for Dutch by the authors of [234].

used by all categories (e.g., both by men and women in the case of gender prediction), usually do not have a strong predictive value for any of the categories. Therefore, the most commonly used approach is to reduce the number of features by applying statistical feature selection methods to extract only those features that show a skewed distribution across the different categories. Wrapper methods, on the contrary, perform a search over the space of all possible subsets of the total feature set, repeatedly training and testing a machine learning algorithm as a subroutine. Hence, each feature subset can be evaluated on a development dataset and the non-discriminative subsets can be excluded from the main series of experiments [67]. However, because testing various feature subsets can become very time-consuming (see [126]) when the number of features is large, which is the case in the datasets used in this dissertation, only single feature selection methods are applied during the experiments.

This study compares the performance of three different single feature selection methods: Document Frequency (DF), Mutual Information (MI) and Chi Square (χ^2). In the following, a brief description of these feature selection metrics is presented.

Chi Square is a common statistical test that measures the divergence between the expected distribution and the observed distribution of a feature for each category. When the divergence is high, the zero hypothesis is rejected, meaning that the frequency of the feature is not independent from the category label. Hence, the higher the divergence, the more discriminative the feature is for a certain category [67]. Where E is the expected and O the observed frequency for all features i in the entire dataset, the Chi Square value can be calculated as follows

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{2.1}$$

Mutual Information provides a measure of the dependency between two random variables [46]. Because it enables researchers to quantify the relevance of a subset of features with regard to the output vector C, the mutual information criterion can be highly useful for performing feature selection. If x and y are two random discrete variables, the Mutual Information is defined as follows [212]:

$$I(x;y) = \sum_{i=1}^{n} \sum_{j=1}^{n} P(x(i), y(j)).log\left(\frac{P(x(i), y(j))}{P(x(i).P(y(j))}\right) \tag{2.2}$$

It is equal to zero if and only if x and y are independent, while higher values indicate a higher dependency (e.g., [118, 181]).

Both methods are included in the experiments, because of their ability to cope with sparse data without making it dense [187]. Finally, although Document Frequency simply measures in how many documents the feature appears, selecting frequent features entails a higher chance that these features will be present in the test data as well, which could be useful when only few features per text sample are available. Hence, this feature selection metric was also included in the experiments.

2.4 Feature Representation

After selecting the features that are most discriminative for the task at hand, each document in the dataset is transformed into a feature vector, in which each vector component refers to the occurrence of a particular feature in the document. The feature vector combined with the document's class label is then referred to as a *document instance*. In this thesis, we compare the performance of numeric vectors, in which each feature is represented by its absolute frequency (Abs.) of occurrence in the document, and three normalised versions, in which the absolute frequency values are scaled between [0, 1]: relative frequencies (Rel.), which are scaled by dividing the absolute frequency of each feature by the total number of features in the document; tf-idf, where the absolute frequencies are multiplied by the inverse of the number of documents in which the feature occurs and, finally, the l_2-norm, in which absolute frequency of each feature is multiplied by the inverse of the square root of the sum of all squared absolute frequencies in the document — also called the Euclidean norm. Additionally, binary vectors, in which the features' presence or the absence is represented as 1 or 0, respectively, are also evaluated. Sparse features are used, which entails that the features with zero value are not (explicitly) included.

2.5 Learning Methods

2.5.1 Supervised Learning

In most text mining studies, the focus is on supervised categorisation of documents, i.e. the task of inferring a model from *pre-labelled* (according to, for example, different authors, genres, topics) natural language text documents. Such approaches typically consist of two phases: the first involves extracting linguistic features that are potentially discriminative for each label from a set of training documents and the second phase builds a machine learning model, so that it is able to attribute new texts to one of the pre-defined labels based on the recurrence of some of the linguistic features that were extracted in the first phase. Next, the resulting classification model is tested on a previously unseen dataset — the test set — containing similar documents (see also Section 2.6). Supervised text categorisation techniques are currently being used in many different contexts, ranging from spam detection, indexing scientific publications and the population of hierarchical catalogues of Internet resources to finding relationships among biomedical entities[2]. To contribute to a systematic analysis of the experimental design, the following set of supervised learning methods was selected, which are all well established in the field of text mining and based upon highly distinctive mathematical algorithms.

Support Vector Machines. SVMs have strong theoretical foundations and have demonstrated their success in a whole range of text mining applications [200]. Given a set of pre-labelled training instances, an SVM algorithm builds a model that can be used to assign one of the labels that were included in the model to new instances — the test data. This model is created by first projecting all instances of the training data into a high-dimensional space. Each instance's location in this space is determined by its vector components (i.e. its features) that are present in the instance. Then, the algorithm computes the optimal separating hyper plane between all instances of the two categories in the task. It does so by maximising the margin between the closest instances of different classes, whereby the instances that lie on the boundaries are the support vectors and the centre of the margin is the optimal separating hyper plane. During the testing phase, each test instance is, in turn, projected into the same high-dimensional space on

[2]An overview of text categorisation techniques and applications can be found in, e.g., [188, 220].

either side of the hyper plane. Finally, the algorithm attributes the test instance to the category that is associated with the side at which it was located. When more than two categories are present in the learning problem at hand, internally the algorithm creates several binary sub-classifiers, adopting either a one-versus-one approach, in which a model is built for each pair of categories, or a one-versus-all approach, in which models are built to distinguish each single category from all remaining categories grouped together. As a result, each categorisation task is treated as a binary problem. Both approaches are followed by a voting phase to determine each instance's final label [45].

Naïve Bayes. Until recently, most applications of supervised text mining were based on Naïve Bayes (NB) methods. Prior literature in pattern recognition and information retrieval which included NB algorithms even dates back for almost sixty years (e.g., [133, 134]). Naïve Bayes classifiers are typically constructed of a family of simple, probabilistic algorithms and work by correlating the presence of each feature in an instance with one of the pre-defined categories. After this training phase, the feature probabilities (or likelihood functions) are used to calculate the probability that a new instance with its particular set of features belongs to one of the pre-defined categories by applying Bayes' rule [18]. Despite their "naive" assumptions, NB classifiers have shown their efficiency in a number of real-life applications such as spam filtering [125].

Random Forests. Random Forests (RF) or Random Decision Forests are an ensemble learning method which includes Decision Trees. Such rule-based classification trees are composed of branches, which represent features, and leaves, which represent decisions. Decision are based on the process of following the branches from the trunk upward until a leaf is reached. Tree learning methods are quite popular in the field of text mining, because they have shown robustness towards the presence of irrelevant features and they produce visually attractive models. However, trees that are grown very deep tend to overfit their training data. Therefore, Random Forests are used to average multiple decision trees that are trained on different subsets of the training data, so that variance can be reduced and a higher performance can be obtained (e.g., [69]).

k-Nearest Neighbor. The k-NN algorithm is a form of Instance-Based or Memory-Based Learning (MBL), which entails that the algorithm stores all training data into memory and uses similarity-based extrapolation during testing. More specifically, it projects all training instances

into vector space and each test instance is assigned the label which has the most representatives within the k nearest neighbors (with k being a pre-defined, positive integer) of that instance (e.g., [4]).

Multi-layer Perceptron. MLP is a class of artificial neural network, which can mainly be distinguished from other methods by its inclusion of one or more non-linear (or hidden) layers between the input and the output layer during classification. The input layer consists of a set of neurons, which represent the features of each instance. Next, each neuron in the hidden layer transforms the values from the previous layer with a weighted linear summation followed by a non-linear activation function. The output layer then analyses the values from the last hidden layer and produces a decision. In most cases, a supervised learning technique called Backpropagation is used during training, which runs a "forward pass" to compute all the activations throughout the neural network and determine the degree in which each node in each layer contributes to any errors in the output of the MLP (e.g., [183, 187]).

2.5.2 Unsupervised Learning

Unsupervised learning, from its part, can be defined as the task of automatically categorising texts based on unlabelled data. In the baseline approach, the similarity between different objects is measured by using one or more similarity functions. With regard to textual data in which the objects can be of different granularities (e.g., documents, paragraphs, sentences or words), clustering methods can be especially useful for e.g., browsing or organising documents and summarising large text corpora [3]. Because the texts provided to the learner are unlabelled, no actual categorisation is performed and, hence, there is no evaluation of the accuracy of the output of the similarity algorithm. Therefore, text clustering is considered an example of unsupervised learning.

This dissertation examines the potential of unsupervised learning techniques to automatically identify different aspects of online child sex offenders' deceptive tactics to seduce their victims (see Chapter 6). More specifically, the following text clustering methods are included in the experiments[3].

[3]A survey of text clustering techniques can be found in e.g., [3].

Hierarchical agglomerative clustering. Hierarchical clustering techniques are designed to build nested clusters by merging or splitting them successively. The hierarchy of the clusters is typically represented in a tree structure or a dendrogram [187]. Agglomerative clustering algorithms perform hierarchic clustering using a bottom-up approach: each instance initiates its own cluster and clusters are merged together using a linkage criteria, such as Ward's algorithm [215].

K-means clustering. The K-means algorithm divides a set of text samples into K disjoint clusters, each described by the centroid of the text samples in the cluster. The algorithm then attempts to select centroids that minimise the within-cluster sum-of-squares (or inertia). It is one of the most commonly used clustering technique for vector data (e.g., [95]).

Finally, *Latent Dirichlet Allocation* (LDA) is a Bayesian probabilistic model, which also assumes a collection of K clusters. In NLP applications the algorithm is often used as a topic model, because it assumes that each document instance is a mixture of a small number of topics and that each word can be clustered into one of these topics (see e.g., [24, 85]).

2.6 Evaluation

During the experiments, we performed five and ten-fold cross validation [219]. In this experimental regime, the available data is randomised and divided into five/ten equally sized folds or partitions. Subsequently, each partition is used four/nine times in training and once in test. This experimental regime provides a more reliable estimation of the performance of a system than when it would only be evaluated on one specific train-test set, because the probability of accidentally selecting an easy or difficult test set is balanced out by the five or ten randomly selected subsets. For each experiment, average scores are reported over all test folds. The scores reported are average precision, recall, F-scores, macro F-scores and accuracy. These are standard evaluation metrics that can be computed based on the number of true positives, true negatives, false positives and false negatives in a confusion matrix. Table 2.6 shows an example confusion matrix for binary classification. In a multi-class set-up, the matrix expands by a row and column for each additional category. When crossing it diagonally, the number of correctly categorised

Table 2.1: Confusion matrix for a binary classification task.

		Predicted Category		
		p	**n**	**Total**
True Category	**p′**	True Positives	False Negatives	P′
	n′	False Positives	True Negatives	N′
	Total	P	N	

documents can be found.

Given a confusion matrix, the recall, precision and F-score for each individual category can be calculated (see [203]). First, the recall score provides information on the number of instances that were successfully retrieved for each class and can be defined as

$$Recall = \frac{TP}{Alltruepositives} \tag{2.3}$$

Secondly, the precision score for each class takes into account all retrieved instances for that class and evaluates how many of them were actually relevant:

$$Precision = \frac{TP}{Allpredictedpositives} \tag{2.4}$$

Next, the F-score for each category is the harmonic mean of the precision and recall score and can be calculated as follows:

$$F = 2x\frac{PrecisionxRecall}{Precision + Recall} \tag{2.5}$$

Additionally, with $L = \{\lambda_j : j = 1 \ldots q\}$ being the set of all categories, the system's performance over L can be evaluated by calculating the macro F-score, which is defined as follows

$$F_{macro} = \frac{1}{q} \sum_{\lambda=1}^{q} F_{\lambda} \tag{2.6}$$

The accuracy is the proportion of true positives and true negatives among the total number of instances in the dataset. As a result, this score is only reported when learning from balanced datasets (see Chapter 4). In other cases, this score could result in misleading conclusions [41].

Finally, based on the classification error, i.e. the percentage of incorrect predictions to the number of predictions made, we evaluate our models' performance by calculating Wilson score intervals [149]. More specifically, where Er is the classification error, $Const$ is a constant value that defines the chosen probability (1.96 or 95%) and n is the number of predictions used to evaluate each model, the Wilson score interval can be defined as

$$Er +/- Const \sqrt{\frac{Er(1-Er)}{n}} \tag{2.7}$$

2.7 Summary

In this chapter, this thesis' approach to the task of automatically detecting adversarial behaviour in online social media communications was introduced. This study adopts a text mining approach, which includes the extraction of information from texts on different linguistic levels and involves the use of statistical and machine learning algorithms for text clustering and classification. Additionally, cross validation is applied, a technique that enables an evaluation the performance for each text mining task on five or ten different partitions of test data, allowing for a more systematic evaluation of the experiments. Finally, a context of the different scores was provided, i.e. precision, recall, F-score, macro F-score, accuracy and Wilson score intervals, which are reported in the following chapters.

CHAPTER 3

THE NETLOG CORPUS: A RESOURCE FOR STUDYING COMPUTER-MEDIATED COMMUNICATIONS

Although in recent years numerous forms of Internet communication, such as email, blogs, chat rooms and social network environments, have emerged, corpora of Internet communications with trustworthy meta-information, such as age and gender or linguistic annotations, are still limited. This chapter introduces a large dataset of Flemish Dutch postings, that was provided to us by Netlog[1]. The NETLOG corpus presented here is a unique resource both for computational (socio)linguistic research in general and for the research purposes discussed in this thesis, because it contains a significant amount of adolescent communications. The lack of such available data has been reported by prior work to be an important drawback for text mining research focusing on applications for protecting children online. Additionally, this chapter aims to provide a better understanding of the level and the role of linguistic noise attested in adolescent computer-mediated communications by performing an systematic study on age and gender related non-standard language variation.

[1] http://nl.netlog.com/

3.1 Introduction

Recent decades have brought a rapid succession of new communication technologies, including text messages and the numerous forms of Internet communication. When compared to earlier (e.g., handwritten) forms of communication, these technology-mediated messages stand out because of an increased level of immediacy. People tend to communicate more and faster, so that their writings become increasingly casual and reminiscent of oral communication. An obvious effect of these recent developments has been the wild proliferation of language variation in written communication, especially affecting surface phenomena such as spelling. While numerous claims have been made about the level of linguistic noise in computer-mediated communications, only few studies have presented a systematic study of the non-standard language varieties in *computer-mediated communications* (CMC). Moreover, due to the lack of available corpus data, none of these works have focused on messages produced by children or adolescents [15].

In the first part of this chapter, the NETLOG corpus is described: a collection of over 1.4 million Flemish Dutch messages, each containing meta-information on the user's profile. The presence of nearly 900,000 online messages produced by children and adolescents under sixteen, makes this corpus a unique resource for both text mining and (socio)linguistic research.

Secondly, we provide a systematic case study of the data: based on a sub-corpus of the NETLOG corpus that was balanced according to age (11 to 49), gender and dialect region (West-Flanders, East-Flanders, Brabant and Limburg), (i) all non-standard words were automatically extracted; (ii) we manually categorised the resulting word list as either typical chatspeak features or belonging to one (or more) of the Flemish dialects or regiolects; and, finally, (iii) we set up a forward stepwise mixed-effects logistic regression analysis [33] to examine the effects of young age and gender on the production of both chatspeak and regiolectal forms.

The methodology presented in this chapter allows for a shift of the research focus from a selection of non-standard linguistic variables, leading to a limited view of non-standard language variation in adolescent CMC (e.g., [211]), to a more systematic approach that incorporates a large number of non-standard words in the selected sub-corpus. Until now, this kind of inclusive approach seems to have been absent in the field [82]. As a result, the work presented here not

only provides a better understanding of the level of linguistic noise attested in Flemish Dutch adolescent CMC, it also leads to a new feature engineering method for extracting non-standard language forms, which is evaluated in the following chapters.

The next section provides an overview of the related work in the field. Section 3.3 describes the NETLOG corpus and illustrates both types of non-standard language use encountered in this dataset. In Section 3.4, we discuss the set-up and the results of the analysis. Section 3.5 concludes this chapter with a discussion of the key research contributions.

3.2 Background and Related Work

3.2.1 Background

In linguistics, non-standard language usage is defined as any language usage that differs from the (un)officially recognized prestige language variant as it is used primarily in written language and formal speech situations [201]. Correlations between the use of non-standard language and sociological variables such as age and gender have already been observed in many spoken discourse studies. In her overview of variationist sociolinguistic research, the author of [198] states that there is a consensus among sociolinguists that of "all the sociolinguistic principles, the clearest and most consistent one is the contrast between women and men", which is defined as the *Gender Effect*. More specifically, (i) in stable sociolinguistic stratification, men tend to use more non-standard forms than women do and (ii) women are usually the innovators in linguistic change [42, 119, 120, 143, 202, 225]. With regard to *Age Grading*, research has shown that people of different ages use speech appropriate to their age group (see e.g., [55, 121, 216]). That is, when a linguistic variety is not part of the standard language, its usage tends to peak during adolescence (i.e. 15–17 year old), "when peer pressure not to conform to society's norms is greatest" [86], while pre- and post-adolescents are found to use these variables less frequently (e.g., [38, 122]). This effect is usually referred to as the *Adolescent Peak Principle*. However, as social pressure increases and the use of standard language becomes more important, for example, for building a career or raising children, people are more inclined to adapt to society's norms. Hence, the use of standard (or prestige) forms tends to peak between the ages of 30 and 55 [198].

Although there is no one-to-one relationship between CMC and spoken discourse (e.g.,[82, 109]), paralinguistic and non-verbal cues, which are absent from the written repertoire, are often compensated by chatspeak features, such as emoticons, character flooding and the use of upper-case to express emphasis (e.g., [49]). Moreover, the loose cross-turn relatedness in multi-participant CMC has not only encouraged language play ([82]), leading to the rise of typical chatspeak abbreviations and concatenations (for example, "bff" (*best friends forever*), "brb" (*be right back*)), it has also led to a trade-off between strictly applying spelling rules and maximizing one's typing speed. For example, errors and typos are seldom corrected and punctuation marks are often left out. Although such features are becoming more common across different cultures and languages, research by [199] showed less convergence of such practices than predicted by [16]. Moreover, they found a tendency to represent "regiolectal spellings" in text messages (SMS) within the US (see also [60, 191]).

3.2.2 Related Research

The study of the variation of linguistic characteristics in CMC according to an authors' age or gender has been a popular subject in both sociolinguistic and computational stylometry studies. At first, most of the studies were based on large collections of weblogs or "blogs", i.e., personal, informal writings listed in reverse chronological order on a blogger's web site [73, 116, 146, 154, 185, 186, 226, 232]. The main advantage of using blog corpora is that blog sites are publicly available and they usually contain information about the blogger's profile. The authors of [114, 116] analysed distributions of non-dictionary words in blogs with regard to age and gender. Although their research focused on automatically predicting age and gender in CMC (see Chapter 4) and not on linguistic variation, the study did report that teenage bloggers tend to use more non-dictionary words than adult bloggers do. Additionally, in their analysis of the number of non-dictionary words used per 1,000 words across the 10s, 20s, 30s and higher age groups, female teen bloggers clearly used more non-dictionary words than the male teens, but the other age groups did not show significant gender differences. However, because other media like online social networks or chat rooms usually do not provide open access to their users' profile, to our knowledge, there is only one study that employs a quantitative approach

to investigate the correlation between the non-standard language usage and profile meta-data in these modes of CMC: in his study on code choice and code-switching in Swiss-German chat rooms, the author of [191] found that younger and older chatters use more dialect than chatters of the middle-aged group, but his study was only based on a list of 70 Standard German words that had a corresponding form in Swiss German. Consequently, he could only investigate about 10 percent of the words in his dataset.

With regard to Flemish spoken discourse, the author of [168] also studied the effects of age, gender and region in a corpus of spoken Dutch transcriptions (the CGN[2], which also includes spontaneous informal speech) on the use of "in-between"-varieties (see Section 3.3.2). His study revealed a correlation between register and region and between register and age, meaning that the dialectic background of the regions is still reflected in informal speech and that Flemish people born in 1940 or later clearly use non-standard language in these situations. No correlation was found between register and gender. However, his study only included a prototypical set of linguistics features, the selection of which was determined by prior knowledge about the central Brabantic regiolect. Moreover, during the compilation of this corpus the informants who participated in the so-called spontaneous conversations all received the explicit request to stick to Standard Dutch, which does not correspond with actual Flemish colloquial speech practices. Furthermore, although there were no researchers present during the conversations, the informants actually had to record them. Both these factors, of course, did not contribute to the spontaneity of the conversations.

What makes the NETLOG corpus interesting for studies on non-standard linguistic variation, is that it was collected without any researchers' interventions and it contains two types of non-standard language use to which the above mentioned Age Grading and Gender Effect could apply: (i) newly incoming non-standard forms — of which (female) adolescents have been shown to be the innovators in spoken discourse (e.g., [198]) — that are characteristic of CMC, and (ii) written representations of regional and dialect forms that are typical for colloquial speech in Flanders (see Section 3.3.2). However, studies on dialect vitality in Flanders (e.g., [208, 210]) have shown that especially younger people tend to use fewer dialect forms than they did a few decades ago.

[2] `http://lands.let.ru.nl/cgn/`

Therefore, in the second part of this chapter the effects of both young age and gender — including potential interactions — on the use of typical chatspeak features and regional speech features are investigated. This case study will include all occurrences of non-standard language use in its analyses, enabling a more systematic study of both feature types. Such a systematic analysis is, to our knowledge, still absent in the field. The set-up and results of the analyses are discussed in Section 3.4. The next section describes the structure and characteristics of the NETLOG corpus.

3.3 The NETLOG Corpus

3.3.1 Structure

Netlog[3] is a Belgian online social networking platform with over 100 million members, utilising over 40 different languages. Members can create a profile page containing blogs, pictures, videos, events and playlists that can be shared with other members. In this section, we introduce the NETLOG corpus: a collection of 1.4 million Flemish Dutch messages, each containing meta-information on the user's profile, which makes this corpus a unique resource for both computational and (socio)linguistic research.

The NETLOG corpus contains 1,465,842 Flemish Dutch postings written by 131,753 different users, with on average 12.4 tokens per message (SD = 38.7). For each posting, information about the age, gender and geographical location of the authors was obtained. Table 3.1 and 3.2 respectively provide an overview of the distribution of the number of users and messages per main Flemish region, gender and age group.

3.3.2 Characteristics of Non-standard Language Variation in Flanders and its Reflection in Chatspeak

All of the social network messages in the present corpus contain varieties of Flemish Dutch. These varieties constitute a vertical and horizontal continuum (see [10]). The so-called horizontal continuum relates to the dialectal and regiolectal geolinguistic variation, whereas the vertical continuum ranges from small-scale local dialects to the standard language, in this case Belgian

[3]In 2018, Netlog is known as Twoo (`https://www.twoo.com/`).

Table 3.1: Number of users per age group, gender and Flemish region in the NETLOG corpus.

Age Group	**West-Flanders**		**East-Flanders**		**Brabant**		**Limburg**		**Total per Gender Group**		**Total per Age Group**
	Male	*Female*	*Male*	*Female*	*Male*	*Female*	*Male*	*Female*	*Male*	*Female*	
11-15	4,048	9,039	5,446	11,058	9,583	19,920	2,138	4,958	21,215	44,975	66,190
16-19	3,139	3,302	3,920	3,705	7,381	6,781	1,539	1,636	15,979	15,424	31,403
20s	1,799	1,238	2,771	1,764	4,207	2,612	1,121	761	9,898	6,375	16,273
30s	740	572	1,021	705	1,392	890	466	339	3,619	2,506	6,125
40s	726	640	924	846	1,226	1,026	520	385	3,396	2,897	6,293
50s+	601	588	783	645	1,066	1,108	359	319	2,809	2,660	5,565
	11,053	15,379	14,865	18,723	24,855	32,337	6,143	8,398			
Total per Category	26,432		33,588		57,192		14,541		56,916	74,837	**131,753**

Table 3.2: Number of messages per age group, gender and Flemish region in the NETLOG corpus.

Age Group	**West-Flanders**		**East-Flanders**		**Brabant**		**Limburg**		**Total per Group**		**Total per Age Group**
	Male	*Female*	*Male*	*Female*	*Male*	*Female*	*Male*	*Female*	*Male*	*Female*	
11-15	37,977	98,218	91,467	173,091	136,943	282,464	17,158	49,806	283,545	603,579	887,124
16-19	22,656	30,387	49,276	41,430	90,662	62,521	9,772	10,171	172,366	144,509	316,875
20s	9,650	5,077	17,312	7,435	27,354	11,626	7,393	2,529	61,709	26,667	88,376
30s	3,736	2,676	7,633	3,986	9,952	3,232	2,982	1,386	24,303	11,280	35,583
40s	5,182	6,687	7,501	9,500	9,463	10,275	4,367	3,631	26,513	30,093	56,606
50s+	8,776	10,472	15,840	9,519	10,184	18,757	3,338	4,392	38,138	43,140	81,278
Total per Category	87,977	153,517	189,029	244,961	284,558	388,875	45,010	71,915	606,574	859,268	**1,465,842**
	241,494		433,990		673,433		116,925				

Standard Dutch, which deviates from Netherlandic Dutch in some minor respects. The diaglossic vertical continuum (see also [9]) is marked by intermediate varieties that are typical of Flemish colloquial speech. Their use seems to be expanding at the expense of the two poles of the continuum [75, 207, 210]. These popular "in-between" or *regiolectal* (rather than dialectal) varieties are still marked by regional differences (e.g., [72]), and can be grouped according to their geographical distribution into four main regiolect areas: Brabant (Antwerp and Flemish-Brabant), West Flanders, East-Flanders and Limburg (see Figure 3.1). Table 3.3 provides an example posting for each region from the NETLOG corpus and its equivalents in Standard Dutch and English. Because these regional non-standard varieties are frequently used in Flemish Netlog users' messages, they are crucial for the present study. However, a complicating factor is that there is no standard spelling for regional features. As a consequence, non-standard forms can be represented by multiple spelling variants. For example, the non-standard Flemish use of "schoon" in the meaning of *beautiful*[4] occurs in the NETLOG corpus as "schoon", "schuun", "sgoown", "sgoon", "skone", "skoon", "skwone", "skwune", "skwoane", "skwunne" and "skwnee". This enormous amount of orthographic variation proves a major challenge in the selection and categorisation of Flemish chatspeak.

Apart from these occurrences of regional non-standard variation, a second type of non-standard language variation, typical of chatspeak in general, is found in the present corpus: Netlog users often omit letters or even entire words or use abbreviations and acronyms in order to maximise their typing speed. In addition, spelling errors are rarely corrected and punctuation marks are often left out. Moreover, to emphasise the content, a flooding of characters or upper-case is often used and (parts of) sentences are concatenated. This is illustrated in Table 3.4. In certain cases, chatspeak features and non-standard regional features are also combined into one single term. For example, the word "wroem" is an abbreviated form of "woaroem", which represents the dialect pronunciation of "waarom" (*why*) in Brabant, while "skwne" is short for "skwunne" or "skwoane", which is the south-eastern West-Flemish variant of "schoon" (*beautiful*).

[4] In Standard Dutch, "schoon" means *clean*.

Figure 3.1: The four main dialect regions in Flanders: West-Flanders, East-Flanders, Brabant and Limburg.

Table 3.3: Examples of non-standard regional varieties in the NETLOG Corpus

Dialect	**Example**	**Dutch**	**English**
West-Flemish	*zitn kik omeki zo verre?*	*Zit ik ineens zo ver?*	*Am I that far suddenly?*
East-Flemish	*est gedon*	*Is het gedaan?*	*Is it over/done?*
Brabant	*wroem niii sgat?*	*Waarom niet, schat?*	*Why not, honey?*
Limburg	*hou gans veel van u*	*Ik hou heel veel van jou.*	*I love you very much.*

Table 3.4: Examples of non-standard chatspeak varieties in the NETLOG Corpus

Type	**Example**	**Dutch**	**English**
Omission	*kbda nimr*	*Ik heb dat niet meer.*	*I don't have it anymore.*
Abbreviations	*wrm, w8*	*waarom, wacht*	*why, wait*
Acronyms	*hvj*	*[Ik] hou van je.*	*[I] love you.*
Character flooding	*keiii mooiii*	*heel mooi*	*very beautiful*
Concatenation	*kweeni*	*Ik weet het niet.*	*I don't know.*

3.4 The Effects of Age and Gender on Non-Standard Linguistic Variation

This section examines the effects of young age and gender on the use of typical chatspeak and regional speech features in a subset of the NETLOG corpus. In the first part of this case study, we describe the extraction and pre-processing of the data sample and discuss the balancing of the data over age, gender and region. Additionally, we explain our operationalisation of non-standard language, together with the parameters we used to categorise each non-standard word into one of both types. Section 3.4.4 provides an overview of the statistical analyses.

3.4.1 Compilation of the NETLOG_SUBSET1

Messages on the Netlog social network can contain multiple quotes from previous posts, of which the correct age, region and gender meta-data are absent in the corpus. Therefore, the first step in pre-processing the data consisted of extracting only the last post of each interaction, of which the required meta-data are available, and saving these as separate files. In the second step of pre-processing the dataset was tokenised, all words were lower-cased and all punctuation marks, emoticons, email addresses, phone numbers and hyperlinks were removed so that only the word forms remained. Additionally, all four or more consecutive identical characters (character flooding) were reduced to three, so that, for example, the tokens “niiice” and “niiiiiic” (*nice*) were considered as the same type[5]. Given that the individual language usage of a Netlog user could bias the results of the analysis if he or she was represented in the dataset by multiple messages, only one post per user was included in the first NETLOG subset. Furthermore, the selected posts contained a minimum of three words. With regard to the age distribution in NETLOG_SUBSET1, the subset was balanced per age year from 11 to 49. No data from users older than 50 were included, because there were not enough data available for each combination of age group and region in the dataset. The smallest group of available users in the corpus was the 34-year-old female group from Limburg with only 27 unique users that had produced a message of at least three words. Therefore, the subset was balanced as follows: 27 postings per age year (11 to 49), gender

[5]Character flooding was not reduced to two, because then the word could become standard Dutch. For example, the intensifier in “zooo slecht" (*sooo bad*) could be wrongly interpreted as the standard form “zoo" (idem).

(male and female) and dialect region (West-Flanders, EastFlanders, Brabant and Limburg). This resulted in 7,952 postings (3,963 female and 3,989 male), which together accounted for 129,358 words. There were 1,998 authors from West-Flanders, 1,998 from East-Flanders, 1,995 authors from Brabant and 1,961 from Limburg.

3.4.2 Identifying Non-standard Language Varieties

The first part of the qualitative analysis presented in this chapter consisted of automatically extracting the number of standard and non-standard words per post. A word was considered non-standard when (a) it was labelled as such by GNU Aspell[6] or (b) it occurred in a list of manually collected non-standard chatspeak words that have a standard Dutch homonym[7]. This resulted in 115,914 (90%) standard and 13,444 non-standard words.

After performing a manual error analysis on 1,000 randomly selected words, the methodology for distinguishing between standard and non-standard words described above yielded an overall accuracy score of 93.8%. False positives for the standard category mostly entailed the non-standard usage of a standard (function) word (e.g., the use of the object pronoun "jou" (*you*) instead of the possessive pronoun "jouw" (*your*) in *"[zou je] me bij jou vrienden brengen"*[8]; homonymy between the non-standard infinitive form and the standard simple past form "zette" (*put*) in *"[zou je] dit op je blog zette"*[9]). Mending these errors, however, would require further research that includes e.g., word sense disambiguation techniques, which is beyond the scope of the present case study. Finally, six personal names were incorrectly labelled as non-standard. The precision, recall and F-scores for each label are illustrated in Table 3.5.

[6]GNU Aspell is an open source spelling checker from which a library for standard Dutch, English and French was imported (`http://aspell.net/`).

[7]This list contains words like "kweet", which in Flemish Dutch CMC is far more likely to be used as a non-standard combination of the clitic "k" + "weet" (*I know*), than as the standard form of the past tense of "kwijten" (*to acquit*).

[8]In English: *[would you] take me to you friends*.

[9]The standard Dutch equivalent would be *"[zou je] dit op je blog zetten]"*. In English: *[would you] put this on your blog*.

Table 3.5: Precision, recall and F-scores (%) for the standard/non-standard feature engineering method.

	Precision	**Recall**	**F-score**
Standard	94.3	95.6	94.9
Non-standard	93.0	91.1	92.1

3.4.3 Categorising the Data

After extracting all non-standard words as was described above, we manually categorised each word as either chatspeak or regional. During this process, all occurrences of spelling deviations and typing errors (i.e. both deliberate deviations from the standard language and unintentional errors (e.g., "vandaga" instead of "vandaag" (*today*), "gecontacteert" instead of "gecontacteerd" (*contacted*)), character flooding (e.g., "waaarom" instead of "waarom" (*why*)) and all abbreviations and other creative adaptations that do not show any regional influence (for example, "bff" (*best friend forever*) and "vr" for "voor" (*for*)) were labelled as chatspeak. In this category, all forms that represented standard spoken Dutch, but were written in a non-standard way were included (e.g., all clitics without regional features like "kheb" for "'k heb" (*I have*), standard Dutch pronunciation assimilations like "feesje" instead of the standard written form "feestje" (*party*) and reduced forms of personal pronouns such as "dachtek" for "dacht ik" (*thought I*)). Additionally, non-standard English forms (e.g., "disign" instead of "design", "wanna" for "want to") were also included in the chatspeak category.

Words in the regional speech category had to display dialect or regiolect features at the level of vocabulary, the representation of vowels and consonants, inflections or conjugations. In other words, although we performed the analysis at the lexeme level, phonological, morphological and lexical deviations from Standard Dutch were taken into account when categorising a word into the regional speech category. In addition, abbreviated forms that showed regional influence in the process of abbreviating were also included. For example, the Standard Dutch word "voor" (*for*) was represented by "vr", "veu" and "vo(e)" in the data: the first abbreviation "vr" does not show any regional influence and was categorised as chatspeak, while the other two varieties lean towards the Brabantic and West-Flemish pronunciation, respectively, and were therefore

Table 3.6: Examples of Flemish regiolect variation in the NETLOG_SUBSET1.

Type	**Example**	**Dutch**	**English**
	begaaie	vuil maken	to smudge
Vocabulary	kozn	neef	nephew
	lekstok	lolly	lollipop
	mokkes	meisjes	girls
	oltid	altijd	always
Vocals	woroem	waarom	why
	zuizu	sowieso	anyway
	weurst	worst	sausage
	skone	schone	beautiful
Consonants	vinne	vinden	to find
	percies	precies	just
	geleje	geleden	ago
	schatteke	schatje	honey
Inflections	dieje	die	that
	broere	broer	brother
	dadde	dat	that
	kgaan	'k ga	I go
Conjugations	hemk	heb 'k	have I
	eje	heb je	have you
	kzin	'k ben	I am

categorised as regional speech. We provide some examples of regional words, their Standard Dutch equivalents and their English translation in Table 3.6. Manually coding the non-standard lexemes in the NETLOG_SUBSET1 resulted in 5,885 regional and 9,146 chatspeak features.

3.4.4 Effects of Age and Gender on Non-standard Language Variation in CMC

In this section, the effects of young age and gender on the distribution of both types of non-standard words (chatspeak and regional speech) in the NETLOG_SUBSET1 are examined. To assess these effects, we employed two separate mixed-effects logistic regression analyses with random by-author intercepts [33].

3.4.4.1 Standard vs. Chatspeak

To allow for non-linearity in the effect of age, we used restricted cubic splines [79], meaning that the age variable was split in a number of intervals with the effect of age on the probability of producing a chatspeak word being fitted as a cubic curve within each interval. At the meeting points of these intervals — the knots — the curves are constrained to have smooth transitions, and beyond the observed values, the independent variable (in this case age) is assumed to be linearly related to the dependent variable (in this case chatspeak probability). Each knot adds a coefficient to the regression model, which can be tested for its statistical significance. There are 123,473 words in this data set: 114,327 (93%) standard words and 9,146 chatspeak words. We clustered the observations in 7,950 authors (3,962 female, 3,988 male). The dataset includes 1,994 authors from Brabant, 1,961 from Limburg, 1,997 from East-Flanders and 1,998 from West-Flanders.

To investigate the age effect, we placed knots at the begin- and endpoint, respectively 11 and 49, and at the ages 27, 33, and 39. The knots at 27, 33 and 39 were added after visual inspection of the mean proportions of non-standard language per age. Additionally, we placed knots at the ages of 15 and 17 to test whether the Adolescent Peak Principle [38, 122] can also be found in the NETLOG corpus. Given that recent studies in Flemish linguistics (e.g., [168, 197, 208, 210]) discuss (and in some cases question) the importance of the Brabant region in the recent development of regiolects that have a wider geographical range than the local dialects do, in this analysis, region was treatment-coded with Brabant as the reference level. The results of the first analysis are summarised in Figure 3.2, which shows the effects of age and gender on the chatspeak word probability. The final model contained significant main fixed

effects for age ($\chi^2(6) = 1038.7, p < .001$), with random intercepts for author ($SD = 0.86$), but not for gender ($\chi^2(1) = 2.58, p = .108$). No significant interactions were found between age and region ($\chi^2(18) = 8.25, p = .975$), age and gender ($\chi^2(6) = 8.79, p = .186$) or gender and region ($\chi^2(3) = 3.75, p = .289$).

As this first analysis shows, age has a significant non-linear effect on the chatspeak word probability. The probability of the production of a chatspeak feature by users from Brabant rose from around 0.15 at 11 to a peak of 0.19 at 15, decreased sharply after that peak until it reached 0.05 at the age of 28. This rise and fall of the chat word probability strongly supports the Adolescent Peak Principle. Although female adolescents are thought to be the innovators of newly incoming non-standard forms — which in this case would be typical chatspeak words like “zjg” for “zie je graag” (*love you*) — there was no interaction between age and gender, which means that the results show no change in the gender effect on the chatspeak word probability during the adolescent peak. Quite unexpectedly, a slight increase was found in the chatspeak word probability between 29 and 33, but after the age of 33, it steadily decreases to reach 0.03 at the age of 41 and remains constant at older ages.

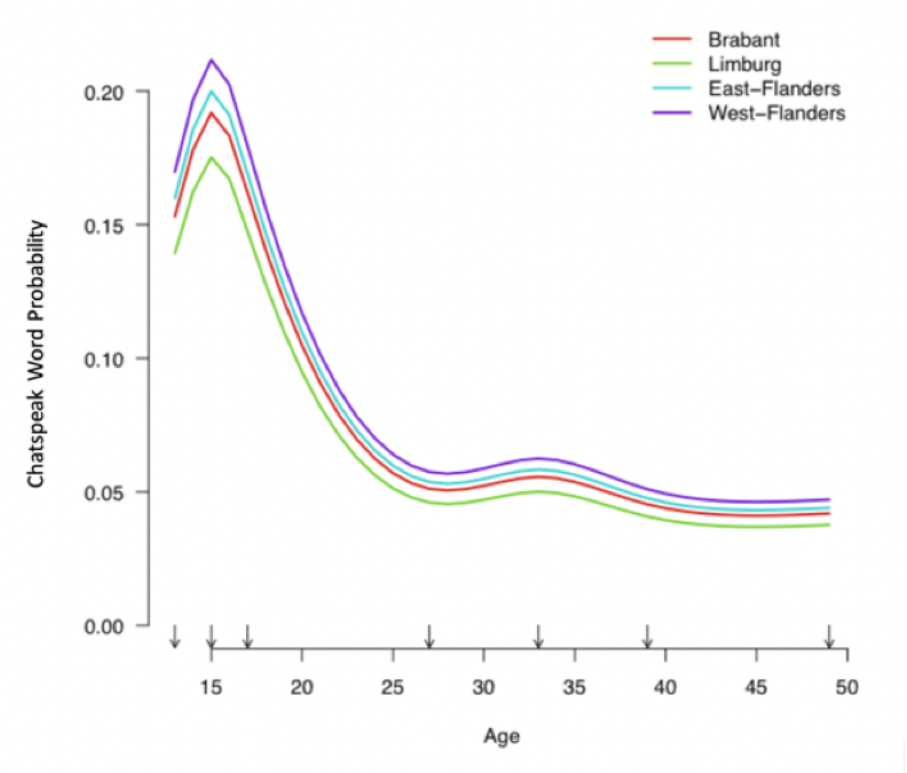

Figure 3.2: Estimated effects of age and region on the probability that chatspeak forms are produced. The curves represent restricted cubic splines with knots placed at the arrows on the x-axis.

3.4.4.2 Standard vs. Regional

Secondly, to assess the effects of age and gender and potential interactions on the probability of a regional word production, we used the same approach as described above. The data set consists of 120,212 words, of which 114,327 are standard words and 5,885 are regional speech features[10]. We clustered these observations in 7,932 authors (3,953 female, 3,979 male). In this analysis, the dataset contains Netlog posts from 1,992 Brabantic, 1,956 Limburgish, 1,988 East-Flemish and 1,996 West-Flemish authors. To capture the non-linearity in the relation between age and dialect word probability, we fitted a restricted cubic spline function with five knots. After visual inspection of the regional word production proportions per author, we placed knots at the endpoints 11 and 49, and at the ages 15, 17, and 33. The knots at 15 and 17 again tested the adolescent peak hypothesis. There were no additional non-linearities to account for ($\chi^2(1) = 0.18, p = .667$).

[10]The standard word observations are the same as the standard words in the chatspeak analysis.

The results of the regional speech use analysis are summarised in Figure 3.3. The second analysis showed significant main fixed effects for age ($\chi^2(4) = 782.36, p < .001$) with random intercepts for author ($SD = 1.30$). Gender again did not show a significant effect on the regional word probability ($\chi^2(1) = 0.64, p = .425$), and there were no significant interactions between gender and age ($\chi^2(4) = 8.89, p = .064$) or gender and region ($\chi^2(3) = 1.55, p = .670$). Although a likelihood ratio test indicated that there was a significant interaction between age and region ($\chi^2(12) = 33.04, p < .001$), none of the interaction coefficients differed significantly from zero (all Wald z p-values $> .16$). With regard to the age variable, as is visualised in Figure 3.3, the probability of producing a regional word for users from Brabant rose from around 0.07 at the age of 11 to a peak of 0.11 at 15, and then decreased smoothly to a value around 0.01 at 49.

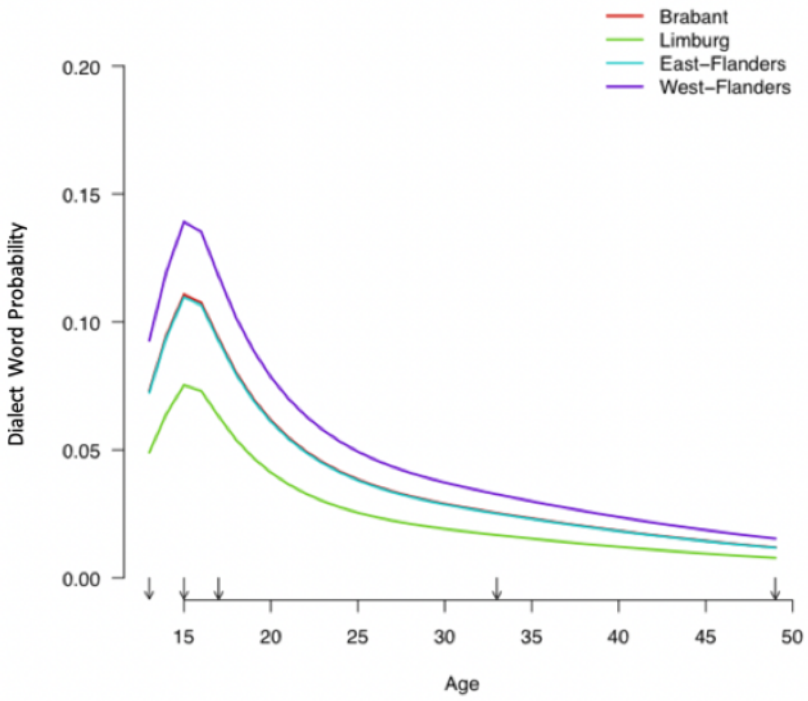

Figure 3.3: Estimated effects of age and region on the probability that regional forms are produced. The curves represent restricted cubic splines with knots placed at the arrows on the x-axis.

3.5 Conclusions and Discussion

In this chapter, we introduced the NETLOG corpus — a considerable collection of very short, Flemish Dutch social network messages, provided by the Belgian social networking platform Netlog that also contains information on the age, gender and location of each user in it. Furthermore, by drawing on a case study of a balanced subset of the NETLOG corpus, we developed a methodology that allows for a more systematic study of the non-standard language variation in adolescent CMC by incorporating a large number of non-standard lexemes in the selected subset. Until now, this kind of inclusive approach seems to have been absent in the research on online social network discourse (see [82]). Secondly, we designed a new feature engineering method to distinguish between standard and non-standard Dutch language use to analyse the widely acknowledged "noisiness" of Flemish social media communications. The research questions examined in this chapter are as follows:

Q1 How linguistically noisy are Flemish social media communications?

Q2 What is the effect of age and gender on both types of non-standard language use (i.e. chatspeak and dialectal forms) in social media communications?

While the results of the analysis presented in this chapter suggest that the NETLOG corpus is indeed linguistically noisy, with 90% of the subset data containing Standard Dutch, applying Natural Language Processing might be more manageable than it is usually perceived to be. This issue is further examined in Chapter 4.

The two forward stepwise mixed-effects logistic regression analyses described in Section 3.4 showed that age had a significant non-linear effect on the production of both chatspeak and regional features, indicating that the principle of Age Grading, that was reported in previous spoken discourse studies of, for example, [55, 86] and [121], is reflected in Flemish social network messages, if only partially, because no data from the older (plus 50) age groups were included in the subset. Additionally, the analyses showed that both the chatspeak word probability and the regional word probability peak between the ages of thirteen and fifteen, which corroborates the presence of the Adolescent Peak Principle. The latter finding is quite remarkable given the

on-going dialect loss processes in Flanders, which mainly affect these younger groups. However, except for West-Flanders, most adolescents produce regional speech with a wide geographical reach rather than small-scale local dialect forms. The latter may not be part of their verbal repertoire any more. In spite of that, they clearly seem to enjoy displaying their (limited) knowledge of regional varieties, the use of which appears to function as a (regional) solidarity marker. Moreover, producing non-standard speech in a written medium dissociates the writing process from orthography and writing standards learnt at school, which makes it extra attractive [191].

Yet, this still does not explain why the peak of the regional word probability starts to decrease after the age of fifteen. Flemish adolescents surely do not possess a more extensive knowledge of regional speech than post-adolescents do and post-adolescents are equally familiar with writing non-standard forms as their younger peers. The presence of an Adolescent Peak in both the chat and dialect word probability could therefore indicate that the use of non-standard words in chatspeak becomes less "cool" during post-adolescence, because it is associated with a younger age group. Likewise, the slight peak in the chat word probability curve around the age of thirty could be linked to the fact that, in Flanders, this age group entails the first generation that acquired "the art" of online chatting. As a result, this could indicate that the use of chatspeak in social media communications is not only attractive for adolescents, but also — to some extent — for people that are currently in their early thirties, because it distinguishes them from the older age groups that did not learn to chat during adolescence. However, these hypotheses require further research to be confirmed. In accordance with [168], no significant effect for gender was found. Additionally, although female adolescents are thought to be the innovators of newly incoming non-standard (in this case chatspeak) forms, the results suggested no change in the gender effect on the probability that chatspeak features are produced during the adolescent peak.

Summarising, the most striking findings of the present study relate, first of all, to the unambiguous illustration of the Adolescent Peak Principle and, secondly, to the intriguing co-occurrence of non-standard regional features and non-standard chatspeak features in written social network postings produced by Flemish adolescents. In the next chapter, we investigate the potential of incorporating these findings to enhance the performance of an automatic user profiling system.

CHAPTER 4

DETECTING AGE AND GENDER IN ONLINE SOCIAL MEDIA: A SCALABILITY STUDY

On most online social networks, it is easy to provide a false name, age and gender in order to hide one's true identity, providing criminals such as child sex offenders with new possibilities to groom their victims. It would therefore be useful if user profiles could be checked on the basis of automatic text analysis and false profiles flagged for monitoring. However, a common characteristic of communication on these digital communities is that it happens via short messages, often using non-standard language variations. Such characteristics make this type of text a challenging genre for natural language processing. This chapter presents a scalability study in which a text mining approach is applied for predicting age group and gender on a subset of the main NETLOG Corpus introduced in Chapter 3, which only contains a single message per user, and examines different strategies for boosting the performance.

4.1 Introduction

Today's generation of people under twenty have barely known a life before social media. For this group, online social networks are interwoven with their everyday life, connecting them to their peers, with hundreds of Facebook friends or Instagram followers to share their thoughts, pictures

and videos with. However, new technology inevitably leads to new types of risks for its users. Prior work by the authors of [177] already demonstrated that digital identities play a crucial role in criminal tactics employed in such digital communities. One offender may hide behind multiple user profiles or a single user profile can be shared by a criminal group when soliciting potential victims. Furthermore, the fluid nature of online identities enables criminals to easily assume a fake identity. As a result, automated techniques for user profile analysis can have a whole range of applications — from cybercrime research, such as detection of child abusive behaviour, fraud detection and identity theft prevention, to social sciences, such as literary science, sociolinguistics and language psychology.

Recent advances in text mining technology have enabled researchers to predict an author's age group and gender in several text genres by automatically analysing the variation of linguistic characteristics. However, in social media, computational linguists are confronted with several issues. First of all, most online social networks do not provide open access to the users' profile data, so it is difficult to collect training data for this task. Furthermore, a recent survey by [15] showed that CMC communications produced by children and teenagers is highly under-represented in existing corpora. Moreover, the few datasets available cannot be used for developing techniques to protect children online, because there either is no specific meta-data available (e.g., the NPS corpus[1]) or the data only contains messages from a very restricted domain (e.g., the Teenage Health Freak website corpus[2]).

Secondly, analysing social media communications typically involves a large number of very short text samples, which inevitably leads to highly sparse data (see Section 4.2.2). So far, the common practice to tackle this problem is to include multiple messages per user in the experiments, resulting in text fragments ranging from 250 to several thousands of words on average per user (see Section 4.2). However, in the NETLOG Corpus far fewer words per user are available (see Chapter 3, Section 3.3).

A third challenge lies within the class imbalance inherent to the task: in most online social networks that are popular amongst children and adolescents (and which can be expected to be

[1]The NPS corpus age range was based on a chat room called "Teen" and contains no detailed information on the actual age of the users represented in the corpus. The corpus can be found here: `http://faculty.nps.edu/cmartell/NPSChat.htm`

[2]`https://www.nottingham.ac.uk/research/groups/cral/projects/thf/thfcorpusoverview.aspx`

targeted by Internet child sex offenders), the number of messages produced by young people highly predominates the number of adult postings. As most machine learning algorithms are designed to optimise the overall accuracy rate, they have been shown to have difficulty identifying documents of the minority class (e.g., [189]).

Fourth, as is demonstrated in the previous chapter, Flemish social media messages contain two types of non-standard language use: newly incoming chatspeak forms and written representations of older regional and dialect forms that are typical for informal conversations in Flanders. Their presence can be expected to pose difficulties for NLP tools that are used to, for example, extract syntactic features from text.

However, for a real-life application — for example, in cybercrime investigations — it is crucial that a user profiling approach is able to achieve a reliable performance despite the challenges mentioned above. This chapter investigates the viability of a text mining approach for performing user profiling on a randomly selected subset of the NETLOG corpus, which only contains a single message per user. As mentioned in Chapter 3, during this work, a dataset of over 1.4 million Flemish Dutch messages was collected, together with meta-information about each user present in the corpus. More importantly, the NETLOG corpus incorporates nearly 900,000 postings produced by adolescents without any domain restrictions, making this a high-value resource for developing child protection applications and computational sociolinguistic research.

Our approach to this computational stylometry task is based on text categorisation, and involves the creation of document representations based on a selected set of (patterns) of linguistic features, feature selection using statistical techniques, and classification using machine learning algorithms (see Chapter 2). Contrary to previous research on age prediction, that mainly focused on predicting age groups (e.g., 10s, 20s, 30s, 40s), the research presented in this chapter focuses on classifying adults versus children. This way, the system should be able to detect adults posing as children and flag their profiles for monitoring. Gender identification is also included in this study, because the vast majority of child sex offenders are male [78, 192] and mismatches between profile gender and predicted gender can therefore also be useful. Furthermore, (predicted) gender can be a helpful information source in constructing more accurate classification models for age. More specifically, the key contributions of this chapter are as follows:

- **a systematic study** of different aspects of methodological design incorporated in a state-of-the-art user profiling approach to assess its robustness to highly sparse, skewed and noisy text data, i.e., performing user profiling "in the wild" (see also [115]);
- **a novel feature engineering method** to distinguish between standard and non-standard language varieties, which have shown to correlate with age and gender (see Chapter 3); and
- **a cross-task classification approach** in which the meta-data for gender is included in the experiments in order to investigate their effect on age prediction.

Finally, this chapter also presents a qualitative analysis of the features that make the user profiling model. Providing such a qualitative analysis of the most discriminative linguistic features enables an evaluation of the scalability of the approach for other researches and allows for methodological reflection towards well established sociolinguistic principles, such as the Adolescent Peak principle, Age Grading and the Gender Effect introduced in the previous chapter. Moreover, a qualitative analysis is typically absent in most text mining studies, as they tend to focus solely on the performance of their user profiling models.

Section 4.2 of this chapter surveys the related work on automatic user profiling, together with a summary of prior research on text categorisation with short texts and the effect of data set size. Section 4.3.1 describes the Netlog subset that is used in the experiments. In Section 4.3 the approach and experimental set-up are discussed. The results of the user profiling experiments can be found in Section 4.4.1 and Section 4.6 provides a qualitative analysis of the features that rendered the best performing models. Finally, Section 4.7 concludes this chapter and summarises its research contributions.

4.2 Related Research

The research presented in this chapter addresses the following key research objectives: the feasibility of detecting age and gender using a text mining approach on the text genre of CMC and the usefulness of the approach when confronted with very short texts, often containing

non-standard language usage. This section describes related research in these areas. A more extensive overview of computational sociolinguistic research can be found in [151].

4.2.1 Automatic Age and Gender Prediction

At first, most studies that involved a computational stylometry approach to automatically predict people's demographics were based on large collections of blogs (e.g., [8, 73, 112, 146, 154, 185, 226, 232]. The main advantage of using blog corpora is that blog sites are publicly available and they usually contain information about the blogger's profile. In one such study, the authors of [186] applied a text categorisation approach to predict gender in a corpus of over 71,000 English blogs. Based on stylistic features (non-dictionary words, parts-of-speech, function words and hyper-links) and content features (content words with the highest Information Gain), they found that "despite the strong stereotypical differences in content between male and female bloggers [. . .], stylistic differences remain more telling than content differences" [186]. However, combining both feature types, they were able to obtain an accuracy of 80.1% when distinguishing between male and female bloggers.

With regard to age prediction, content words proved to be slightly more useful than the style-based features, but again combining them rendered the best results: 10s were distinguishable from 30s with accuracy above 96% and differentiating between 10s and 20s was achieved with an accuracy of 87.3%. However, many 30s were wrongly classified as 20s, which rendered an overall accuracy of 76.2%. This resulted in an F-score of 0.86 for the 10s, 0.75 for the 20s and 0.52 for the 30s category[3]. The authors of [226] were the first to include "non-traditional" features in their experiments, such as background colour, word fonts and cases, punctuation marks and emoticons. When combining these non-traditional features with bag-of-word features, their system achieved an F-score of 0.68 based on a corpus of 75,000 English blog entries authored by 3,000 individual bloggers. Interesting to see was that removing stop words actually decreased the performance of their system to 0.64, which is consistent with previous sociolinguistic studies that attested gender differences in the use of highly frequent word classes such as pronouns, articles and prepositions (e.g., [22, 108, 136, 137, 147, 150]). Similar results were found for age: based on

[3]These scores were calculated based on the confusion matrix in the paper.

the same corpus as was described in [186], the authors of [114] showed that language usage in blogs correlates with age: pronouns and the use of both assent and negation become scarcer with age, while prepositions and determiners become more frequent. Their system yielded an accuracy of 76.1% for the three-way classification problem of attributing blogs to one of three age groups: 13–17, 23–27 or 33–47 (majority baseline = 42.7%) by combining style- and content-based features and 80.5% for predicting gender. The authors of [73] further expanded the research of [186] by adding non-dictionary words and the average sentence length as features. Furthermore, the stylistic difference in usage of non-dictionary words combined with content words allowed to predict the age group (10s, 20s, 30s or higher) with an accuracy of 80.3% and gender with an accuracy of 89.2%. The average sentence length, however, did not correlate significantly with age or gender. Additionally, [176] found that female authors were more likely to use emoticons, ellipses, character flooding, repeated exclamation marks, puzzled punctuation (i.e., combinations of "?" and "!"), the abbreviation "omg" (*oh my god*), and transcriptions of back-channels like "ah", "hmm", "ugh", and "grr". Affirmations like "yeah" and "yea" were the only preferences that were attributed to males. These latter features are called — not quite accurately — "sociolinguistic features" in e.g., [176]. Finally, a number of other, non-textual features have been suggested for age and gender prediction, such as the number of friends and followers [5, 176] and posted images [226]. The current study is limited to linguistic features extracted from each message.

More recently, a number of studies were based on a corpus of Twitter (e.g., [5, 13, 21, 66, 152, 176]) and other social network data (see e.g., the author profiling tasks at PAN 2013, 2014 and 2015 [173–175]). Although the amount of available data on Twitter is expanding massively, profile data is often absent, which requires additional techniques to acquire such meta-data. Contrary to blogs, tweets are typically very short, containing a maximum of 140 characters. However, most studies tend to combine multiple messages per user and show very similar results to previous studies on weblog data. The best results for gender prediction were achieved by the authors of [13], whose system achieved an accuracy score of 88.0% based on over 600 tweets per user. When predicting age on a corpus of 200 Dutch tweets per user, [152] were able to reach a 0.76 F-score when distinguishing between users younger than 20, between 20 and 40 years old and older than 40. Binary age prediction (adults versus adolescents), as examined in this chapter,

was first performed by the authors of [65], who investigated the performance of shallow textual features (e.g., character counts), language models and non-textual information (e.g., number of friends) when identifying bloggers under and over 18. However, their classifiers only yielded slightly better results than their majority baseline. Finally, the authors of [177] presented a set of tools for predicting age and gender in a forensic context. By including POS, semantic and BOW features in a hierarchic classification system, their hierarchical, binary age prediction model yields probabilities that a user belongs to a specific age band (11–18 or over 18, followed by a breakdown of the probabilities for 11–14; 15–18; 19–49; 50+; etc.), resulting in a 72.15% recall and 72.24% precision for distinguishing between children and adults.

Aside from investigating which feature types are most effective for predicting profile information, the authors of [232] contributed to the field by comparing different data representation methods, feature selection methods and machine learning algorithms for gender prediction in 3,226 blogs (52% female), which contained about 400 words on average. They also included 20 semantic labels (e.g., "conversation", "family") as features in their instances, which were based on lists of words appearing in a similar context (e.g., "tell", "talk", "ask" belonged to the "conversation" label). Together with these word factor analysis features, they included word unigrams, POS tags and average word and sentence length in their experiments, but did not compare the results of these feature types individually. Their best prediction accuracy of 72.1% was achieved by using Information Gain as feature selection criterion, and Support Vector Machines (linear kernel) as machine learning algorithm. Based on a corpus of 3,100 English blogs with an average post length of 250 words for men and 330 words for women, [146] investigated which feature selection methods were most suitable for their type of data. Their ensemble feature selection method (EFS) improved the accuracy scores on gender attribution significantly compared to single selection metrics, such as Information Gain and Chi Square, by about 6-10%, resulting in a best accuracy score of 88.6%. Although this EFS method showed promising results, its application in age and/or gender attribution remains limited to [146]. The reason for this could be that building a new classifier for each subset remains very time-consuming when working with a large number of features.

Although these studies show promising results for both age and gender prediction in social

media communications, all of the previously mentioned studies included text fragments ranging from 250 to several thousands of words on average per user. However, when looking at a recent study by [34], these high results are subject to scalability issues when the models are applied on shorter text fragments: although their implementation of the Balanced Winnow algorithm yielded an accuracy score of 75.5% for predicting gender when using multiple tweets per user in their instances, the performance decreased significantly to 66.5% when using only a single tweet per user. Therefore, this chapter investigates the scalability of a text mining approach when confronted with limited data availability. The next section describes prior work related to short text categorisation.

4.2.2 Short Text Classification

With the recent proliferation of online communications and e-commerce, processing short text samples is becoming increasingly important in many Internet applications. However, contrary to more traditional forms of written texts, such web-based datasets are typically less topic-focused, linguistically noisier, imbalanced and much shorter (they contain a few sentences at most). Generally, the features of short text are as follows:

- **sparseness**: most documents only contain a small percentage of the total number of features present in the dataset. Because of their limited length, they provide great challenges for standard text mining approaches that rely on word frequencies, word co-occurrences or shared context to determine the similarity between documents;
- **skewed big-data**: in many cases the focus lies on detecting the minority class in a big-data set-up and, hence, the number of useful instances is limited.

A range of different techniques have been suggested to address these issues. The authors of [194] presented a summary of the current approaches to short text classification. They distinguish three broad types of research: semantic approaches, semi-supervised classification and ensemble methods. Semantic approaches are based on enriching short text samples by extrapolating external taxonomies, such as Wikipedia or Probase (e.g.[14, 88, 214]), Wordnet (e.g., [26, 88]) or including implicit information produced by search engines (e.g., [26, 103]). However, although

such approaches eliminate the problem of data sparseness, they also produce a large number of overzealous features and, hence, they are said to add to the problem of dimensionality [195]. Additionally, combining smaller amounts of labelled data with larger, unlabelled datasets (i.e., semi-supervised classification) was reported to yield a higher performance when assigning topic to technical papers [231] and classifying document titles [193].

Within a computational stylometry framework, the effect of data size has not been researched in much detail, since most of those studies tend to focus on long texts or several short texts per author [131]. The author of [35] regarded 10,000 words per author to be a reliable minimum for an authorial set. In a study on short texts by the Brontë sisters, the authors of [84] found that using multiple short texts can reduce the problem of dealing with short texts, even when "short" means only 200 words per author. Furthermore, the authors of [184] stated that 5,000 words in training could be considered a minimum requirement. However, when reducing the number of words per text fragment to 100 words, the author of [132] reported a dramatic decrease of the performance for authorship attribution.

This chapter provides a systematic study of each aspect of the experimental design which could affect the performance of a text mining approach to automatic user profiling on short, non-standard social network communications. Additionally, different strategies are investigated to boost the performance of the resulting models for age and gender prediction. The next section elaborates on these different strategies.

4.3 Experimental Set-up

Contrary to prior research that included a large number of messages per user in their experiments (see Section 4.2), the first part of this chapter investigates the feasibility to predict age group and gender on the message level. More specifically, it examines four stages in experimental design that are inherent to a text mining experiment and can be expected to have a high impact on the ability of the approach to deal with highly sparse, linguistically noisy datasets: feature types, feature selection, feature representation and machine learning. So far, such a systematic comparison of decisions in experimental design that affect the performance of the approach when

confronted with these challenges has been absent in the field.

Based on the results of this systematic study, the second part presents three strategies to boost the performance for automatic user profiling using only a single message per user. Because, in the context of a cybercrime investigation, a conjunction of evidence with external content, which was not produced by the suspect[4] would be out of the question, only the following strategies are explored:

1. **a feature union approach** in which different feature types are extracted in parallel and the resulting vectors are concatenated into larger vectors to create complex models;

2. **a balancing strategy** in which all experiments were performed a second time: (a) balancing the dataset in each training partition, but maintaining the original skewed datasets in the test partitions of the ten-fold cross validation scheme; and (b) balancing the entire dataset to enable a valid comparison to previous work in this area; and

3. **a cross-task classification approach** in which the meta-data for gender is included in the experiments in order to investigate their effect on age prediction.

To this end, we compiled a new subset[5] of the NETLOG corpus, which contained only one message per user. Section 4.3.1 discusses the compilation of this subset and its pre-processing procedure.

4.3.1 Data

As we discussed in the previous chapter, Netlog messages can contain multiple quotes from previous postings. Because the corpus does not contain the users' profile information of these additional texts, the first step in pre-processing consisted of extracting the last posting of each interaction and saving these as separate documents. For example, from the postings that are displayed in Figure 4.1 we only extracted the answer from "xElke_" to the message written by "_Zahraah__X".

[4]For example, by including information from Wikipedia, Wordnet or Internet search engines as is typically done in both the semantic and semi-supervised classification approaches described in Section 4.2.2.

[5]This subset contains a different selection of Netlog postings than the one described in Chapter 3, Section 3.4.1.

The *Diagnostic and Statistical Manual of Mental Disorders* [6] defines paedophilia as a sexual paraphilia where the offender has to be at least 16 years old and at least five years older than the victim; the victim being not older than 12 or 13. However, because the legal minimum age for sexual interactions is set at 16, the following age classes were created based on the demographics meta-data: MIN16 (up to 15 years old) and PLUS20 (20 and older). Furthermore, because research in psychiatry and behavioural sciences has shown that most individuals that engage in paedophilia are male (e.g., [78, 192]), the following complex classes were also included: PLUS20_MALE and PLUS20_FEMALE. Furthermore, because the illegality of, for example, relationships between an 18 year old and a 15 year old, is often very difficult to determine without a thorough police investigation, no data from users between 16 and 19 were included in the experiments. Next, for each user in the NETLOG corpus, one posting was randomly selected. This second NETLOG subset will be referred to as NETLOG_SUBSET2.

Finally, we tokenised all messages. Because of the presence of non-standard language forms (see Chapter 3, Section 3.3), the tokenisation process consisted of splitting up each post in a list of words, emoticons and punctuation marks, but concatenated forms (e.g., "iloveyouso") and incorrectly spaced compounds (e.g., "kei lang" instead of "keilang" (*very long*)) were left unchanged. Additionally, all tokens were lower-cased and email addresses, phone numbers, hyper-links and xml-tags that refer to text formatting (i.e., bold, italics, underlined) were converted to a single type each (e.g., "[$email$]", "[$hyperlink$]", "[$formatting$]"). Finally, character and punctuation flooding were reduced to three characters, so that "byeeeeee" and "byeee" (*bye*) were considered the same type (see also Chapter 3, Section 3.4.1). All original forms were saved in separate documents together with their adjusted forms so they could still be used for feature engineering.

Figure 4.1: Example of a Netlog message containing one quote and an answer to that quote.

4.3.2 Feature Types

As explained in Chapter 2 (Section 2.2), most text mining studies include (combinations of) lexical, character and syntactic features in their experiments. In this study, three new feature types are introduced, which together are referred to as *sociolinguistic* features. This section describes the NLP procedure for extracting these feature types.

Lexical features. Unlike some previous studies on age or gender prediction (see Section 4.2), we did not create a dictionary of hand-picked features that are likely to distinguish between different age or gender categories. Instead, we applied a data-oriented approach by extracting all token unigram features (i.e., a bag-of-words model), including punctuation marks, emoticons, pre-processed email addresses, hyperlinks and text formatting tags, allowing for a more complete picture of the age and gender related linguistic variation in the data. To enable a similar data-oriented approach for content and function words individually, a list of function words is required. Although such a list in itself is limited in Standard Dutch, when analysing Flemish chatspeak the approach is confronted with numerous linguistic variations: aside from concatenations and abbreviations such as "kheb" for "ik heb" (*I have*) and "gn" for "geen" (*no*), the NETLOG corpus also exhibits a wide variety of regiolectal function words that deviate from Standard Dutch on the phonological or morphological level (e.g., "veur" instead of "voor" (*for*), "werre" instead of "worden" (*to become*)) or even on the lexical level (e.g., "golle" in Brabant versus "junder" in West-Flanders, meaning "jullie" (*you*)). Therefore, we extended the list of standard function words with the most commonly used non-standard function words from each region, which we extrapolated manually from the qualitative analysis described in Chapter 3 (Section 3.4.2). The content word features included all other words together with the emoticons, but without punctuation marks. Additionally, to include context information in the experiments, we extracted token bi-, trigrams and skip-grams[6].

Character features. As character n-grams have been shown to be useful for tracing stylometric evidence beyond topic and genre [185] and proven to be reliable when dealing with limited data [131, 159], we included character bi-, tri- and tetragrams in the user profiling experiments.

[6]Skip-grams are a variety of n-grams in which the features are not necessarily consecutive, but may leave gaps that are "skipped" over [77].

Syntactic features. We generated Part-of-speech tags using Frog [27], a Dutch morpho-syntactic analyser and dependency parser. However, because this NLP tool was trained on Standard Dutch, messages containing non-standard language forms proved to be problematic. This is illustrated in Table 4.1 with an example of the Frog output of a Netlog posting containing non-standard language use and the output of its equivalent in Standard Dutch. However, given that the same tagging errors are made both when analysing the training and test folds, it is possible that these errors will only marginally affect the features' performance.

Semantic features. This feature type has been suggested by [167] for predicting age and gender. More specifically, they used the Linguistic Inquiry and Word Count (LIWC) word category lexicon [166], which includes sixty-four different categories of language, among which topical categories, such as family, affect, occupation and body. For example, words like "father", "mother", "brother" and "sister" all belong to the "family" category. We used a semi-automatic Dutch translation of the LIWC2015 lexicon [165] as described in [209] to extract semantic features during the experiments.

Table 4.1: Frog analysis of a Netlog posting and its Standard Dutch Equivalent. The labels marked in blue are correct.

Example[7]	**Frog Output**	**Dutch**	**Frog Output**
kzeg	N(soort,ev,basis,onz,stan)	*ik*	VNW(pers,pron,nomin,vol,1,ev)
		zeg	WW(pv,tgw,ev)
wel	BW( )	*wel*	BW( )
dak	N(soort,ev,basis,onz,stan)	*dat*	VG(onder)
		ik	VNW(pers,pron,nomin,vol,1,ev)
ziek	ADJ(vrij,basis,zonder)	*ziek*	ADJ(vrij,basis,zonder)
zn	VNW(bez,det,stan,red,3,ev,prenom,zonder,agr)	*ben*	WW(pv,tgw,ev)

4.3.2.1 Extracting Sociolinguistic Features

SNS features. The statistical analyses described in Chapter 3 demonstrates that sociolinguistic principles such as Age Grading and the Adolescent Peak Principle also apply on CMC as can be found in the NETLOG corpus. Therefore, for each token, we extracted information on its "standard-

ness": a word was labelled non-standard when (a) it was analysed as such by GNU Aspell or (b) it occurred in our list of manually collected non-standard chatspeak words that have a standard Dutch homonym (see Chapter 3, Section 3.4.2). Next, for each message in the NETLOG subset, each word was represented as either a "[standard_content]", "[non-standard_content]", "[standard_function]", "[non-standard_function]" feature based on the output of this operationalisation, together with the manually extended list of function words described above. When evaluating the first 1,001 words from the first test fold of the NETLOG_SUBSET2 dataset (see Section 4.4.1), the methodology for extracting SNS features yielded an overall accuracy of 92.7% with a classification error of 7.3% +/- 1.6%. Moreover, the results showed that this method also performed well for three of the four labels individually: with an F-score of 98.1%, the best results were achieved when detecting standard Dutch function words. With regard to detecting standard and non-standard content words, this resulted in an F-score of 92.1% and 92.5%, respectively. For the non-standard function words, however, the approach only achieved a 77.2% F-score. The precision, recall and F-scores for each label are stated in Table 4.2.

Table 4.2: Precision, recall and F-scores (%) for SNS feature engineering method.

Features	**Prec.**	**Rec.**	**F-score**
[non-standard_function]	74.2	80.3	77.2
[standard_function]	96.3	100.0	98.1
[standard_content]	92.1	91.9	92.0
[non-standard_content]	94.4	90.7	92.5

When performing a manual error analysis on the data itself, it appeared that there was some minor confusion between non-standard words and non-standard function words, which mostly can be linked back to either spelling variations of the non-standard function words that were not included in the manually composed list, such as "daz" for "da's" (*that's*), or to abbreviated words that could also refer to a non-standard word that is not a function word (e.g., "nr" can be an abbreviated form of both "naar" (*to*) and "nummer" (*number*)). Nonetheless, these words were still correctly labelled as non-standard language use. However, there are also nine cases in which a function word was incorrectly labelled as standard Dutch. After a closer inspection of the data,

Table 4.3: Confusion matrix (absolute values) for the sociolinguistic features' engineering method.

CM (Abs.)	[NS_function]	[NS_content]	[S_function]	[S_content]
[NS_function]	**49**	3	5	4
[NS_content]	5	**303**	1	25
[S_function]	0	0	**237**	0
[S_content]	12	15	3	**339**

these errors mostly entailed non-standard usage of a standard (function) word. Similar results were found for most of the words that were incorrectly labelled as standard (e.g., homonymy between the non-standard infinitive form and the standard simple past form "zette" (*put*) and vice versa, even across different languages (e.g., homonymy between the standard English and the non-standard Dutch "my" in "my abv: brintleey!"[8]) (see also Chapter 3, Section 3.4.2). Finally, six personal names were incorrectly labelled as non-standard, because they did not occur in the list of named entities.

Paralinguistic features. In chatspeak, paralinguistic and non-verbal cues — that are present in spoken discourse, but absent in the formal written repertoire — are often compensated by other linguistic features, such as emoticons, character flooding and the use of upper-case and punctuation flooding to express emphasis [49]. Hence, information on the presence of these paralinguistic features were also included in the document representations. More specifically, prior to pre-processing, the occurrence of character and punctuation flooding were extracted and represented as "[char_flood]" and "[punct_flood]" features, and the occurrence of non-standard capitalisation as "[char_upper]". Additionally, emoticons and other combinations of characters and punctuation (e.g., "< 3" representing the shape of a heart) were represented as "[punct_play]" features. Together with the standard/non-standard operationalisation described above, these feature are referred to as *sociolinguistic features*.

4.3.3 Feature Selection and Representation

Various techniques have been suggested to represent a feature's presence and frequency in a document instance. This thesis investigates whether any of the commonly used methods described

[8]The standard Dutch equivalent would be "mijn" instead of "my" here. "abv" is the abbreviated form of "allerbeste vriendin" (*very best girlfriend*).

in Chapter 2 shows a considerably higher performance when dealing with a high-dimensional, sparse dataset containing short text samples. Within the experiments, five frequently used feature representation methods are analysed: binary representation (Bin.), absolute frequencies (Abs.), relative frequencies (Rel.), tf-idf and $l2$-norm rescaling (l_2). Additionally, a comparative analysis is performed between four different feature selection methods: Document Frequency (DF), Chi Square (χ^2) and Mutual Information (MI), which are used to select the 10,000; 50,000; and 75,000 most discriminative features.

4.3.4 Machine Learning

For the purpose of this study, a range of different machine learning methods were run on the NETLOG_SUBSET2 dataset. More specifically, for classification, we examined the performance of the following algorithms[9]:

- *Support Vector Classification*. Because SVMs have demonstrated robustness to high-dimensional data and imbalanced text mining problems (e.g., [188, 189]), two SVM implementations were included in the experiments: C-Support Vector Classification (with RBF kernel, $c = 2048.0$) (C-SVC) and a linear SVM with Stochastic Gradient Descent (SGD)[10] training. The first is standard in a text mining approach. The latter was included because of its increasing popularity in "big data" ML applications.

- *Naïve Bayes (NB)*. The Naïve Bayes algorithm is popularly used for traditional text classification purposes and has shown tolerance to missing values [117]. For classification, a Multinomial NB algorithm was used, because the algorithm outperformed Gaussian and Bernouilli NB algorithms during the preliminary experiments.

- *k-Nearest Neighbor (k-NN)*. The k-Nearest Neighbor algorithm was suggested by [131] for performing authorship attribution "in the wild" and has demonstrated robustness towards overfitting [117]. The k parameter was set to 5. Neighbor weights were assigned proportionally to the inverse distance from each test instance.

[9] All classifiers were trained using the LibShortText [228] or Scikit-learn [187] machine learning packages.

[10] When applying SGD training, the gradient of the loss is estimated per instance and the model is updated along the way with the learning rate [187].

- *Random Forest (RF)*. As the Random Forest algorithm has shown tolerance to missing values and irrelevant attributes in prior text categorisation research [117], which could be relevant to the task at hand, it was included in the experiments. The number of trees in the forest is set to 10; the Gini impurity function [30] was used to split a decision tree.

- *Neural Network (NN)*. A Multi-layer Perceptron classifier was trained using Backpropagation (see Chapter 2, Section 2.5). The log-loss function was optimised using stochastic gradient descent as proposed in [110] ('adam'), which has shown efficiency towards large datasets [187]. Also, MLPs have been used to derive robust features for e.g., speaker recognition [81].

The parameters reported above were optimised during the preliminary experiments on a validation sample of the training data.

4.3.5 Evaluation

To obtain a reliable estimation of the single classifiers' performance, we applied ten-fold cross validation (see Chapter 2, Section 2.6). In this experimental regime, we randomised and divided the available data into ten equally sized partitions. Subsequently, we used each partition nine times in training and once in test.

This section provided an overview of the experimental set-up that was adopted to systematically compare different aspects of experimental design of a text mining approach when confronted with the challenges mentioned earlier. In the next section, we discuss the results for age group identification and gender prediction on the message level and provide insight in the effect of the methodological decisions made.

4.4 Part I: Effect of Experimental Design on Single Classification Models for User Profiling

Social network messages differ from other CMC text genres, such as e-mails or blogs, in several aspects. The length of each instance is usually much shorter, their vocabulary and grammatical structure are often non-standard and the distribution of lengths is very similar: the average post length in the NETLOG_SUBSET2 is 12.7 tokens, with 90% of the posts containing maximum 30 tokens. In this section, we examine the behaviour of a text mining approach to automatically identify Netlog users' age group and gender under the complex conditions of short social network postings containing non-standard language use. To evaluate different aspects of methodological design, first, we conducted a series of 576 experiments, in which we examined the performance of three feature selection techniques (DF, χ_2 and MI), four feature representation methods (tf-idf, bin., Abs. and $l2$-norm.) and six different machine learning approaches (C-SVC, SGD, NB, k-NN, RF and NN) for both age and gender prediction on the message-level, using the highly skewed NETLOG_SUBSET2 in a ten-fold cross validation set-up. When analysing the effect of each aspect on the performance of the model, the other factors were kept constant to allow for a valid comparison. Next, we used the best performing model to compare the performance of lexical, character, syntactic, semantic and sociolinguistic features for both tasks.

4.4.1 Age Group Identification

We calculated random baselines by generating predictions according to the following strategies: (a) stratified, (b) uniform and (c) majority. *Stratified* random classification (RC) generates predictions by respecting the training data's class (imbalanced) distribution; *uniform* RC yields predictions uniformly at random and a *majority* RC strategy labels each instance with the most frequent category in the training set. The random baselines according to each strategy are displayed in Table 4.4. Given the objective of detecting adults posing as adolescents, we focus on the performance of each combination for the adult class.

In the first series of experiments we only included token unigrams (BOW) as features. As can be observed in Table 4.5, all five machine learning approaches were able to improve upon

Table 4.4: Random baselines for age group identification in the NETLOG_SUBSET2 dataset.

Strategy	PLUS20			MIN16		
	Prec.	*Rec.*	*F-sc.*	*Prec.*	*Rec.*	*F-sc.*
Stratified	18.2	18.5	18.4	83.0	82.8	82.9
Uniform	16.9	49.0	25.1	82.5	50.0	62.2
Majority	0.0	0.0	0.0	82.8	100.0	90.6

the baseline performance with regard to the minority class (PLUS20) — despite the high class imbalance and the short text samples attested in the dataset. The highest F-score for both the PLUS20 and the MIN16 category — hence, the best overall performance — was obtained when using a Multinomial Naïve Bayes classifier. When representing the 50,000 most discriminative features with their absolute, tf-idf or binary values in each document instance, the NB classifier achieved a 66.5% F-score for identifying adults (DF) and a 94.2% F-score for detecting children between 11 and 15 years old (χ_2) based on only a single message per user. Both implementations of SVMs (C-SVC and SGD) and the NN classifier produced slightly lower best F-scores for the older age group (64.6%, 63.9% and 63.6%, respectively), but the k-Nearest Neighbor and the Random Forest algorithms yielded a notably lower performance when identifying the PLUS20 category (52.1% and 52.9%). With regard to the younger age group, all algorithms were able to improve upon the random baselines, demonstrating very similar results: the best performance was achieved by the NB classifier with an F-score of 94.2%, followed by both SVM algorithms (93.5%), the NN classifier (93.3%), the RF algorithm (92.6%) and k-NN (92.2%).

With regard to feature selection, almost every machine learning approach benefited from reducing its dataset's dimensionality to its 50,000 most discriminative features, except for the k-NN classifier. Applying Document frequency led to the best F-scores for the older age group when using the Naïve Bayes and the Random Forest Classifier, while performing feature selection based on the Chi Square method improved the performance of the C-SVC and the NN algorithms for the same category. With reference to the younger age category, the best results were achieved in almost all cases by applying Chi Square. When analysing the different feature representation methods, it is clear that the performance varies according to the machine learning algorithm

they are combined with. More specifically, the NB classifier produced similar results when incorporating tf-idf, binary and absolute feature values and marginally lower results when $l2$-normalisation was applied. Conversely, the other ML algorithms tended to yield slightly better results when using the latter representation method. Table 4.5 provides an overview of the best performing combination of feature selection and representation methods per machine learning algorithm. The best performing model showed a classification error of 17.5% +/- 0.2%.

Table 4.5: Results (%) for the best performing combinations of feature selection and representation methods per machine algorithm for age group identification in the NETLOG_SUBSET2 dataset.

			Scores					
ML	**Feat. Rep.**	**Feat. Sel.**	PLUS20			MIN16		
			Prec.	*Rec.*	*F-sc.*	*Prec.*	*Rec.*	*F-sc.*
C-SVC	l2	χ_2 (50K)	72.3	58.3	64.6	91.7	95.4	93.5
SGD	Bin./l2	MI (75K)/χ_2 (10K)	70.4	58.6	63.9	90.4	97.0	93.5
NB	Bin.	DF (50K)/χ_2 (50K)	77.1	58.4	**66.5**	90.9	97.7	**94.2**
NN	Tf-idf/l2	χ_2 (75K)/χ_2 (10K)	68.8	59.2	63.6	91.0	95.8	93.3
RF	Tf-idf/Abs.	DF (10K)/χ_2 (50K)	66.0	44.1	52.9	88.1	97.6	92.6
k-NN	Tf-idf	All feats.	67.0	42.6	52.1	88.9	95.6	92.2

After this first series of experiments in which different aspects of experimental design were examined for their robustness towards short social network communications, a second series of experiments was performed to investigate whether other feature types could improve upon the performance of the best performing model from the first series. More specifically, the performance of three traditional feature types was examined, which could potentially be informative for predicting age in Flemish chatspeak: lexical (token unigrams, content and function words), character n-gram (bi-, tri- and tetragrams) and syntactic features (POS uni-, bi- and trigrams). Aside from these types, this second series of experiments also investigated whether training on a new type of features, i.e., sociolinguistic features (see Section 4.3.2), could potentially enhance the performance of the profiling system. For reasons of comparability, each feature was represented by its binary value and feature selection was performed using DF.

When analysing the results of the single feature types, the BOW features outperformed all other feature types when identifying the adult class, yielding a precision of 77.1%, a recall of 58.4% and an F-score of 66.5%. Stemming each token in the model increased the precision to

94.0%, but had a negative effect on recall and F-score (28.9% and 44.2%, respectively). Focusing on either content or function words only, the NB classifier was not able to produce better results than the bag-of-words model. Additionally, including context information in the experiments (i.e., by using token n-grams and skip-grams) did not enhance the performance. With regard to the non-lexical feature types, character tetragrams achieved the best results, yielding a higher precision of 81.8%, but a lower recall and F-score of 54.4% and 65.3%, respectively. Syntactic features (POS) performed very poorly and training on semantic (LIWC) or sociolinguistic features as a single feature type resulted in the classifier attributing the majority label to each instance. Finally, the best performing single feature types to identify the MIN16 category were the character tetragrams and BOW features. An overview of the results is displayed in Figure 4.2 and 4.3 in Section 4.5.1.

4.4.2 Gender Detection

We performed a similar series of experiments to investigate the effect of methodological design on a text mining approach when distinguishing between male and female Netlog users in the NETLOG_SUBSET2 dataset. Random baselines for gender detection can be found in Table 4.6. Additionally, because research in psychiatry and behavioural sciences has shown that most individuals that engage in paedophilia are male (e.g., [78, 192]), we focus on the performance of each combination for the male category.

Table 4.6: Random baselines for gender detection in the NETLOG_SUBSET2 dataset.

Strategy	MALE			FEMALE		
	Prec.	*Rec.*	*F-sc.*	*Prec.*	*Rec.*	*F-sc.*
Stratified	26.8	27.3	27.1	75.6	75.2	75.4
Uniform	25.2	51.0	33.7	75.2	49.5	59.7
Majority	0.0	0.0	0.0	75.0	100.0	85.7

As can be seen in Table 4.7, the results from the gender classification experiments are similar to those of the age prediction models with regard to experimental design: the best F-scores for both gender categories were achieved by the Naïve Bayes classifier when including the 50,000 most

frequent token unigrams (DF). Representing features with their absolute values resulted in a best F-score of 38.8% for the MALE class, while using tf-idf feature values led to 86.3% for the FEMALE category. All models were able to improve upon the random baselines when identifying the minority class, but only the NB and the SGD classifiers produced better results than the majority random baseline when detecting female Netlog users. Again, both SVM models (C-SVC and SGD), together with the NN classifier achieved slightly lower results for gender prediction than the NB classifier, while the performance of the Random Forest and the k-Nearest Neighbor algorithms was considerably lower for both categories. Furthermore, applying Document Frequency for performing feature selection yielded the best results when detecting the minority class. The best combinations of feature selection and representation methods per classifier are shown in Table 4.7. The best performing model showed a classification error of 23.5% +/- 1.0%.

Table 4.7: Results (%) for the best performing combinations of feature selection and representation methods per machine algorithm for gender detection in the NETLOG_SUBSET2 dataset.

			Scores					
ML	**Feat. Rep.**	**Feat. Sel.**	MALE			FEMALE		
			Prec.	*Rec.*	*F-sc.*	*Prec.*	*Rec.*	*F-sc.*
C-SVC	Bin./l2	χ_2 (50K/10K)	43.3	29.5	35.1	77.5	97.0	86.1
SGD	Tf-idf/l2	DF/χ_2 (10K)	42.9	29.6	35.0	77.9	93.4	85.0
NB	Abs./Tf-idf	DF (50K)	55.9	29.8	**38.8**	77.0	98.2	**86.3**
NN	Bin.	DF(10K)/χ_2(50K)	36.9	17.2	23.5	75.6	98.3	85.5
RF	Abs./l2	MI(50K/75K)	45.8	15.3	22.9	76.4	95.7	85.0
k-NN	Bin./l2	χ_2 (75K/10K)	40.7	30.5	34.9	78.4	91.7	84.5

Next, to investigate which feature types were most informative when predicting gender, a NB classification model was trained on other lexical, character, syntactic, semantic and sociolinguistic features, which were extracted by applying the Document Frequency selection metric and selecting the 50,000 most discriminative features using absolute feature values. Contrary to the age detection experiments, character n-gram features outperformed all other single feature types for the minority class, with character tetragrams achieving a best F-score of 44.1%. Of the other feature types, only the bag-of-words model and the BOW skip-grams were able to improve upon the baseline performance for the MALE category, yielding a 38.8% and 38.3% F-score, respectively. Removing content or function words or stemming did not increase the performance. Although

the classification models trained on semantic (LIWC), syntactic (POS) or sociolinguistic features were not reduced to a majority label classifier (see Section 4.4.1), they scored considerably lower than the lexical and character *n*-gram features. With regard to identifying the female Netlog users in the NETLOG_SUBSET2 dataset, all single feature types produced considerably higher F-scores compared to the stratified and uniform random baselines. However, aside from the BOW model described earlier, they were not able to improve upon the majority baseline for the FEMALE category. The results of each feature type can be found in Figure 4.4 and 4.5 in Section 4.5.1.

4.5 Part II: Boosting Strategies

Based on the results of the systematic study of different aspects of methodological design presented in the previous section, this section examines three different strategies to boost the performance for automatic user profiling using only a single message per user: a feature union approach (Section 4.5.1), a balancing strategy (Section 4.5.2) and a cross-task classification approach (Section 4.5.3).

4.5.1 Combining Features into Complex Models

Because the experiments that were based on token unigram and character 4-gram features showed the best results for both age groups, in the next step we combined both these types with other single feature types. Interestingly, as can be seen in Figure 4.2 and 4.3 there was only one combination that was able to top the performance of the BOW features for both categories: when merging character 4-grams with sociolinguistic features, the NB classifier achieved a 83.3% precision, a 56.2% recall and a 67.1% F-score for the adult class and a 91.5% precision, a 97.7% recall and a 94.5% F-score for the adolescent age group. Threefold combinations of single feature types did not produce higher results for either category.

Similar to the feature union experiments for age group detection described above, the best single feature types were merged with other types to examine which combinations could boost the performance of the gender classifier. Again, the best results were achieved when combining character tetragrams with the sociolinguistic features introduced in this chapter, resulting

in a 47.8% precision, 45.1% recall and 46.6% F-score for the male category. With regard to identifying females, the BOW model still outperformed all other combinations. An overview of the performance of the different (combinations of) feature types is presented in Figure 4.4 and 4.5.

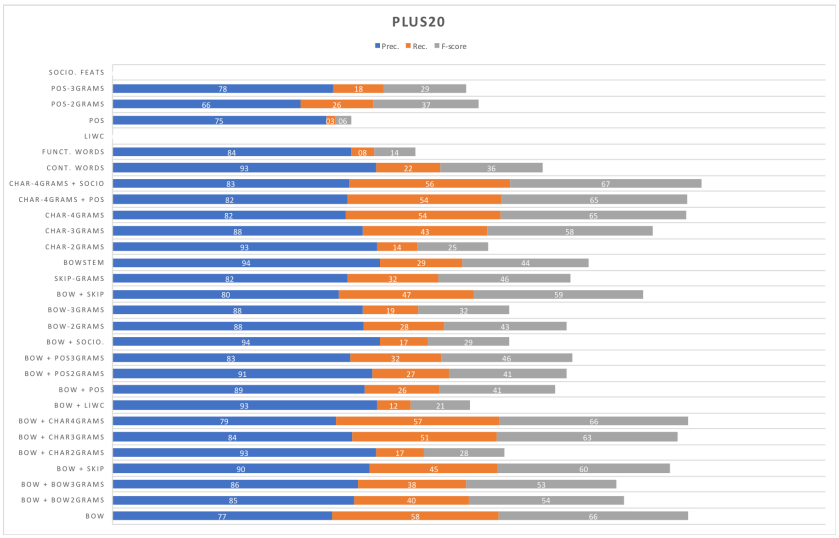

Figure 4.2: Precision, recall and F-score for age prediction per feature type (combination) for the PLUS20 class.

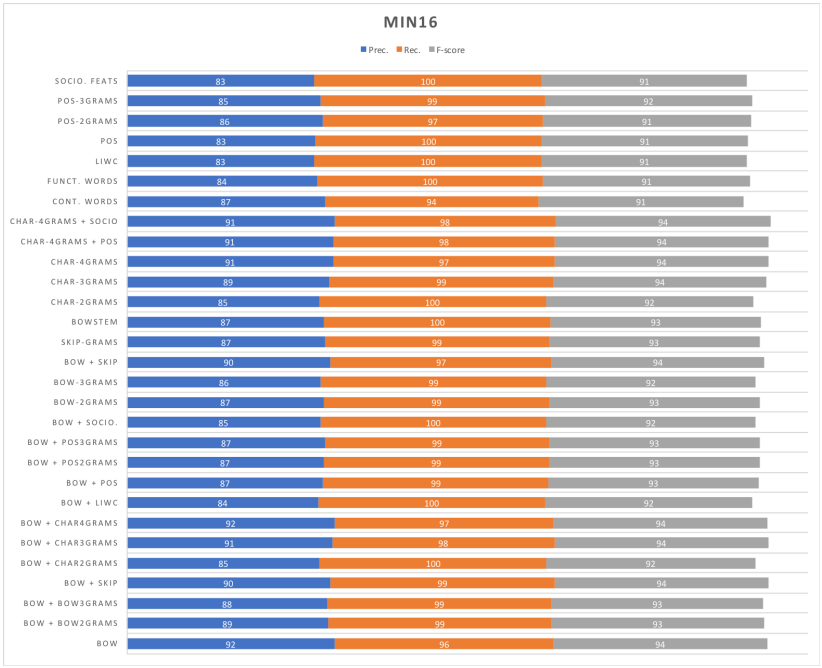

Figure 4.3: Precision, recall and F-score for age prediction per feature type (combination) for the MIN16 class.

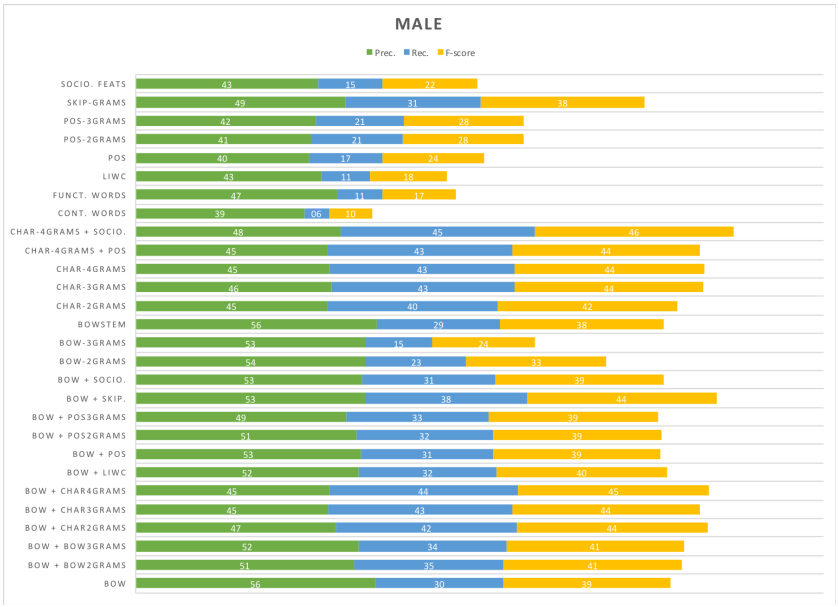

Figure 4.4: Precision, recall and F-score for gender prediction per feature type (combination) for the MALE class.

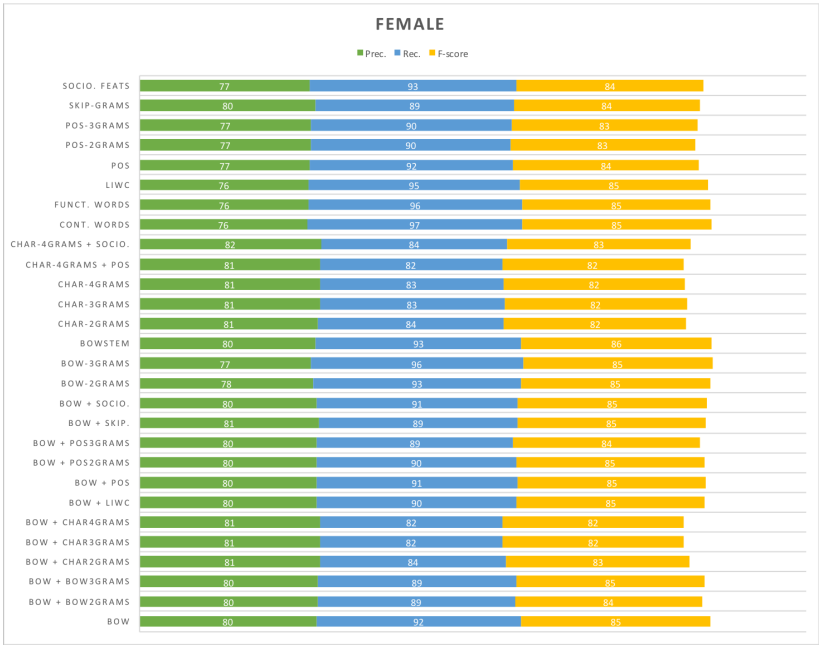

Figure 4.5: Precision, recall and F-score for gender prediction per feature type (combination) for the FEMALE class.

4.5.2 Balancing the Data

To create a good reflection of reality, up until this point, a highly skewed data distribution was adopted during each learning experiment. The results discussed above showed that it is feasible to improve upon the baseline performance for the minority classes that are the focus of this study (i.e., male adults), based on a heavily imbalanced dataset containing only a single message per user. The series of experiments described in this section investigated whether balancing the data in train while maintaining the original skewed data distribution in test could increase the performance for these categories. Furthermore, experiments on the completely balanced NETLOG_SUBSET2 dataset were also included to enable a useful comparative analysis with regard to prior studies in automatic user profiling (see Section 4.2.1).

The best results for both types of balanced data experiments were again obtained by the Multinomial Naïve Bayes classifier, using a combination of the character 4-grams and sociolinguistic features. Balancing the dataset in each training partition only, led to a considerably higher recall score for the minority classes when predicting age and gender compared to the imbalanced data experiments described above, but the precision and F-score decreased in for both the PLUS20 and the MALE category.

Finally, regarding the completely balanced learning experiments, this study's findings show that adding the sociolinguistic features (which were newly introduced in Section 4.3.2) to the more traditional character n-gram features produced a higher accuracy of 84.1% compared to prior work for age prediction on blogs (e.g., [73, 116, 186]) and social network postings ([152, 204]) which all incorporated multiple messages per user in their experiments. With regard to gender prediction, the Naïve Bayes classifier yielded slightly lower results (61.0% accuracy) than reported by [34], who set up a learning experiment in which a single tweet per user was incorporated. However, the authors of [34] did not include adolescent messages in their dataset.

4.5.3 Combining Age and Gender Prediction

Finally, the third part of this section investigates whether gender meta-data can be a helpful information source in constructing more accurate classifiers for age group detection.

In view of a comparative analysis to the results of the binary classification experiment for age

(see Section 4.4.1), which are referred to as BIN_EXP., in this section, three different approaches of including the meta-data for gender are examined in order to investigate their effect on age group prediction. Given that one of the key objectives of this thesis is to investigate the feasibility of detecting adults posing as adolescents in social network environments, during these experiments the focus lies on the scores for the adult class.

In the first experiment (EXP_1) the data was balanced according to both age and gender in each training partition of the NETLOG_SUBSET2 dataset. Next, the Multinomial Naïve Bayes classifier was retrained, extracting the 50,000 most frequent features (DF) and using binary feature values (see Section 4.4.1). Compared to the results of the imbalanced data experiments, both the recall and F-score for the PLUS20 category improved from 56.2% to 81.9% and from 67.1 to 69.8%, respectively. The precision, however, decreased from 83.3% to 60.8%.

In the second experiment (EXP_2) a three-way NB model was trained on the original, imbalanced dataset, which included the MIN16 category and two complex classes for the older class in which gender was included, namely PLUS20_MALE and PLUS20_FEMALE. Subsequently, the complex classes were reduced to PLUS20 in the classifier's output in order to examine whether the extra gender information the classifier had acquired during training would lead to a better age prediction on the binary test sets. Although the recall dropped slightly to 80.9%, the results of EXP_2 showed an improvement upon the EXP_1 approach with regard to the precision (67.7%) and the F-score (73.7%).

The third experiment (EXP_3) consisted of including gender as an additional feature in every instance of the original NETLOG_SUBSET2 data (i.e., a feature union approach). Again the results improved upon those of BIN. and EXP_1, but not upon those of EXP_2, resulting in a precision of 65.8%, a recall of 81.4% and an f-score of 72.8% for the adult class. In Table 4.8 an overview is provided of the results for the three additional experiments compared to those of the binary age classification experiments (BIN_EXP.). In EXP_2, the model showed a classification error of 16.2% +/- 0.2%.

Table 4.8: Results for age prediction when including gender meta-data: EXP_1 (data balanced according to age group and gender in train), EXP_2 (3 classes in train, 2 in test) and EXP_3 (gender as feature).

Scores (%)	**Class**	**Bin.**	**Including Gender**		
			Exp_1	**Exp_2**	**Exp_3**
Precision	Min16	91.5	**96.0**	95.9	95.9
	Plus20	**83.3**	60.8	67.7	65.8
Recall	Min16	**97.7**	89.0	92.0	91.2
	Plus20	56.2	**81.9**	80.9	81.4
F-score	Min16	**94.5**	92.4	93.9	93.5
	Plus20	67.1	69.8	**73.7**	72.8

4.6 Qualitative Analysis of Predictive Features

In recent literature, word clouds have become very popular for visualising, comparing and summarising texts. However, most of these tools do not allow for visualising the relationship between different features. Therefore, to enable a qualitative analysis of the 100 most informative features per category, the results were not only scaled according to their score, but semantically related words were also placed close to each other. To produce such semantic words clouds, the following analyses were performed[11]:

- *Feature Extraction*: we extracted character 4-gram features together with sociolinguistic features.
- *Ranking*: the features were ranked according to relative importance. We used tf-idf as ranking function.
- *Computing Term Similarity*: we applied Cosine Similarity to calculate similarity values between features.

[11]The qualitative analyses presented in this section were produced by using the SWCV toolkit (`https://github.com/spupyrev/swcv`).

- Finally, an edge-weighted graph was created in which vertices correspond to the terms and weights correspond to the calculated similarities. For each feature, an axis-aligned rectangle is created with height proportional to its rank. We applied the Star Forest approximation algorithm (as described in [20]) for constructing contact representations of the graph.

This yielded a set of non-overlapping positions for the rectangles with semantically related words placed close to each other [17].

As is illustrated in Figures 4.6 to 4.9, the results of the experiments show that teenage social media messages differ considerably from that of the older age groups — in their choice of words, their syntax and their use of non-standard forms, intensifiers and paralinguistic cues, which is in line with prior sociolinguistic studies describing the Adolescent Peak Principle ((e.g., [38, 122], see also Chapter 3, Section 3.2.1).

Contrary to the results of the statistical analyses that included both the chat and dialect word probability described in Chapter 3 and the work of [168], who studied the effects of age, gender and region in the CGN corpus on the use of "in-between"-varieties and found no correlation between register and gender, the experiments presented above did show a positive effect on the gender prediction experiments when including sociolinguistic features. This can be explained by the presence of emoticons, character flooding and the use of upper-case and punctuation flooding, which were integrated in the sociolinguistic features, but were excluded from the standard vs. non-standard operationalisation adopted in the statistical analyses in Chapter 3. Such features are used to compensate paralinguistic and non-verbal cues that can be found in spoken discourse, but, up until the rise of CMC, were absent in the written repertoire. These findings are in line with prior work by e.g., [42, 119, 120, 143, 202, 225], who found that women are usually the innovators in linguistic change. Because no correlation was found between gender and the use of dialect/regiolect or chatspeak features, it can be concluded that the Gender Effect is only partially present in the Flemish NETLOG Corpus.

Figure 4.6: Semantic Word Cloud Visualisation of the 100 most discriminative features for the MIN16_MALE category.

Figure 4.7: Semantic Word Cloud Visualisation of the 100 most discriminative features for the MIN16_FEMALE category.

Figure 4.8: Semantic Word Cloud Visualisation of the 100 most discriminative features for the PLUS20_MALE category.

Figure 4.9: Semantic Word Cloud Visualisation of the 100 most discriminative features for the PLUS20_FEMALE category.

4.7 Conclusions

Most text mining studies in computational stylometry are able to successfully identify the age and gender group of a given text when relying on large amounts of training data per user. However, when applying user profiling "in the wild", i.e., in online social media, often only limited and linguistically noisy data is available. Additionally, the task typically implies a big-data set-up, in which the classes are often unequally represented. Hence, an approach is required which scales towards such a challenging text type. The experiments described in the first part of this chapter were designed to investigate the feasibility of predicting age and gender under the complex conditions of a large, skewed dataset containing highly sparse, noisy text samples.

Experimental design, while one of the most critical factors when designing a text mining

approach for any real-life application, is also one of the least studied. Therefore, in this chapter, we systematically analysed the performance of different methodological aspects, such as feature extraction, selection, representation and machine learning when performing user profiling on only a single message per user. Additionally, we examined three different strategies for their ability to boost the performance, while keeping each user's original message intact, which is essential when developing a methodology that can be used to support cybercrime investigations. In this chapter, the following research questions were addressed:

Q3 Can including linguistic noise contribute to a higher performance of a text mining approach designed to perform user profiling in online social media?

Q4 What is the effect of different feature selection, feature representation, document representation and machine learning techniques on the ability of a text mining approach to deal with highly sparse data?

Q5 Which feature types show more robustness under such complex conditions?

Q6 Which techniques can be used to increase the performance of the approach?

Despite the challenging characteristics of this text genre for natural language processing, this study showed that it is feasible to improve upon random baseline performance for both age and gender classification using highly limited data sets of on average 12.7 tokens (i.e., words, (Netlog) emoticons and punctuation marks) per instance. The best results for both tasks were achieved when developing complex models in which character n-grams were combined with a new feature type, which was inspired by sociolinguistic principles, such as the Gender Effect and the Adolescent Peak Principle, established by prior work on spoken discourse corpora. Additionally, the findings described in this chapter indicate that character features are not only robust when dealing with limited data [131, 159], when combined with information about linguistic noise (i.e., the sociolinguistic features), they also showed to be useful for tracing stylometric evidence "in the wild".

When training on the 50,000 most informative features (DF), the Naïve Bayes classifier showed the best results for classifying adults versus adolescents, yielding an F-score of 66.5% for

the PLUS20 and a 94.2% for the MIN16 category, and for performing gender detection, achieving an F-score of 38.8% for the MALE and 86.3% for the FEMALE class. Balancing the data in each training partition produced higher recall scores for the minority classes in both tasks, but also resulted in a lower precision and F-scores. Since gender could be a helpful information source in constructing a more accurate classifier for age, in this chapter three different approaches to including the meta-data for gender were examined. First, we balanced the training data according to both age and gender. Secondly the NB classifier was trained on complex classes, which included both age and gender information and reduced to two classes in post-processing, which only related to age, and finally a binary classification experiment was performed with gender as additional feature in each instance. Although all three approaches showed improvement compared to the results of the initial age detection experiments, the best F-score of 73.7% for the adult class was achieved by training on an imbalanced dataset in which the users were labelled as MIN16, PLUS20_FEMALE or PLUS20_MALE and reducing the complex classes to PLUS20 in post-processing.

A question that remains, however, is how these models will perform when the adults are posing as an adolescent, a tactic often used by online child sex offenders to establish contact with their victims and gain their trust [177]. In order to answer this question, the next chapter focuses on an adversarial stylometry approach in which the best performing models described in this chapter are crash-tested on adversarial text samples.

CHAPTER 5

AN ADVERSARIAL STYLOMETRY STUDY TO DETECTING FALSE USER PROFILES

"amaai zo mooi meiske da gij zijt, ik wou dak kik u was se!"

"zin om een geheim met me te delen?"[1]

The experiments described in the previous chapter show that a text mining approach can be used in a social network environment for distinguishing adolescents from the older age categories — despite the challenging characteristics that are innate to the genre. Up to this point, we assumed that the majority of users in the NETLOG corpus did not have the intention to disguise or misrepresent their profile information. However, a key issue when designing a user profiling system is whether the approach will remain reliable when it is confronted with false user profiles. With this research question, this thesis addresses a relatively new sub-domain of computational stylometry, namely adversarial stylometry, which investigates whether writing style is (partially) unconscious or whether it can be obfuscated or imitated.

[1] These messages were taken from the PREDATOR corpus (see Chapter 6). In English: *wow, you are such a beautiful girl, I wish I were you! Want to share a secret with me?*

5.1 Introduction

In recent years, Darknets and other environments offering anonymity are becoming increasingly popular among child sex offenders with a high degree of computer literacy and forensic awareness. Additionally, the emergence of online communities of people who share a sexual preference for children are enhancing the "normalisation" of the crime. Not only do they offer "moral support", such online hubs also provide them with technical, security and seduction tips [96]. Although none of such anonymisation techniques (e.g., the Tor service [52, 178]) is entirely bulletproof, they can easily complicate or even block cybercrime investigations. In such cases, the communications produced on social media platforms can be one of few clues to an offender's identity [180].

The posts quoted above were taken from the PREDATOR corpus (see Chapter 6) and were written by a child sex offender who claimed to be a sixteen-year-old girl. Hence, these posts illustrate one of the key issues both moderators and law enforcement agencies are faced with today, namely that such posts only become suspicious in the case of an adult writing them on a young girl's page. Unfortunately, it is virtually impossible to manually check the vast amount of communications online in order to tackle the risk of children being solicited by online child sex offenders.

Up to this point, we assumed that the majority of users in the Netlog data had not set up a false user profile. However, the main question when designing a user profiling system that could be used to support cybercrime investigations, is if the methodology will remain useful when it is confronted with adversarial text samples. Initial work in this area is not very encouraging. The authors of [31], for example, demonstrated that even state-of-the-art authorship identification methods could be reduced to random behaviour when they are confronted with passages that include obfuscation, in which people attempt to hide their identity, or imitation, during which people try to mimic an other author's writing style [32]. However, like previous age prediction studies (see Chapter 4), most adversarial stylometry research often included relatively large samples of texts in their experiments (see Section 5.2). As a result, it could very well be that the number of clues revealing the authors' own writing style were too limited to stand out between the majority of deceptive features in these larger text samples.

Within the context of the present study, it is our hypothesis that adults — especially when they are in touch with young people's culture — will be able to adopt a younger chatspeak style, enabling them to deceive children and perhaps even other adults, but that their adult stylistic fingerprint (or rather "writing print") will unwittingly flow through in (parts of) their communications. Hence, this chapter investigates whether a *message-based* approach, in which predictions are made on the level of the individual post and aggregated to the user level in post-processing, enables the system to identify these clues of deception more accurately than the traditional *user-based* approach that renders predictions directly on the user level. We apply both approaches (a) in the context of automatically detecting people trying to imitate the writing style of another demographic group and (b) on the text genre of short, online social media communications. So far, both these applications have remained unexplored in the field.

To investigate this hypothesis, we created two additional datasets: to enable a comparative analysis between the post- and user-level experiments, we compiled the NETLOG_SUBSET3, that was based on 13K different users and contains 15 posts per user; additionally, in order to test the performance of both approaches on adults creating a fake child persona, we created the VOLUNTEER corpus, which consists of 24 chatspeak conversations between a child and an adult volunteer posing as an adolescent. This set-up also allowed for an evaluation of both approaches across different online social media (i.e., social network versus chat room).

The rest of this chapter is structured as follows. Section 5.3 describes the compilation of the NETLOG_SUBSET3 and the VOLUNTEER datasets. The experimental set-up and the results of the experiments are described in Section 5.4. The key findings of this chapter are summarised and discussed in Section 5.6. The next section provides an overview of the related research.

5.2 Related Work

As is mentioned in Chapter 1, stylometry is based on the assumption that every individual has a unique writing style and, as a result, an author can be distinguished from other authors by measuring specific properties of his or her writings. However, most stylometric research is also based on the assumption that authors do not attempt to disguise their linguistic writing style. The

author of [98] discusses the importance of determining the robustness of an authorship attribution system when it is confronted with deception. However, so far, research into this issue has been limited. The authors of [102] were the first to explore the possibility to computationally obfuscate the (most likely) author of the disputed Federalist Papers (see e.g., [155]). They attempted to hide the author's identity by neutralising 14 of the most informative words per thousand words in the texts. Yet, the obfuscation was successfully detected by a technique called *unmasking*, which was proposed by [116]: using a series of SVM classifiers to iteratively remove the features that received the highest weight from the SVM's during training, they found that, when comparing two texts that were written by two different authors, the accuracy score slowly declined during the iterations. However, when comparing two texts that were written by the same author of which one was computationally modified, as in [102], they attested a steep drop of the accuracy. This drop is explained by the fact that when comparing between texts that are written by the same author, the number of highly discriminative features is limited. Hence, when building a degradation curve, iteratively removing these features typically results in sudden drops in accuracy. However, when comparing texts that are written by different authors, the number of discriminative features is larger, resulting in a more steadily declining accuracy when iteratively removing subsets of these features (see also [106]).

Contrary to these studies, however, initial work by [32] showed that including obfuscation passages written by humans resulted in a devastating effect on the robustness of most state-of-the-art authorship attribution methods. Moreover, in an extended study on the English Brennan-Greenstadt corpus, which included original writings, obfuscated, and imitation passages of 45 different authors, the work of [1] stated that including obfuscation passages resulted in a decrease of the precision for authorship attribution from over 80% to less than 10% when training on data from forty different authors. With regard to detecting imitations of literary writings (or pastiches), the authors of [53] reported that using frequency rankings of stop words as features showed promising results when trying to distinguish between the Romanian novelist Caragiale's writings and authors that had attempted to imitate his writing style after his death. However, research by [1] reported a precision of less than 5% when including imitation passages from the English Brennan-Greenstadt corpus.

Although the work of [100, 101] confirmed the fact that identifying the author of such deceptively written texts is extremely difficult, both [1] and [99] found that the authors' intent to deceive or hide their identity is detectable. On the one hand, the authors of [1] reported that, in imitated passages, the usage of personal pronouns and particles increased, while the usage of adjectives decreased. They also noticed an increased use of existential "there", adverbs, particles and personal pronouns in obfuscated passages, but a decrease in the usage of nouns and wh-pronouns. Finally, they noticed that authors tend to "dumb down" their writing style by using shorter sentences, simpler words with less syllables, lower readability scores and higher readability ease and that changing function words seemed to be an important way to obfuscate a text. On the other hand, using the Java Graphical Authorship Attribution Program (JGAAP) software package, [99] was able to identify five of six "deceptive" documents (83%) and 22 out of 28 "honest" writings (79%) in the Brennan-Greenstadt corpus. He concluded that the attempts of people to write "differently" could be fit into a recognisable and distinctive stylistic pattern. Yet, the research described above still shows a number of limitations in the context of this study: (a) some of the mentioned studies were performed on computationally modified data instead of on deceptive texts written by (untrained) human beings; (b) the studies that did include obfuscation or imitation passages written by humans were all performed within the framework of authorship attribution studies and did not investigate the feasibility of masking demographical characteristics; and, finally, (c) all studies were performed on a minimum of 500 words per author and only included formal text genres. Hence, the question addressed in this chapter is whether a human being is able to modify his or her writing style in such a way that a user profiling system attributes the imitated label — in our case that of the adolescent — instead of the true adult label in the context of short, informal social media communications. The next section describes the different datasets used in this chapter.

5.3 Data

5.3.1 The NETLOG_SUBSET3

To enable a valid comparison between the message- and user-level profiling experiments, we extracted all MIN16 and PLUS20 users who had produced at least fifteen posts from the NETLOG Corpus. As in real life, this resulted in a heavily skewed dataset with regard to the age and region categories: 78% of the users are adolescents (11 to 15 years old), 46% originate from the Brabant area (Antwerp or Flemish-Brabant), while 20%, 26% and 10% of the users are located in West-Flemish, East-Flemish and Limburg users, respectively. The user distribution according to gender is highly skewed in the younger age category, with 61% of the total number of users belonging to the MIN16_FEMALE category. With regard to the older age category, as can be seen in Table 5.1, the NETLOG_SUBSET3 is more balanced.

Based on this subset of Netlog users, we randomly selected 15 messages per user (resulting in a total of 204,060 postings) and created two datasets for our experiments: the POST_LEVEL and the USER_LEVEL datasets, in which the document representations per user contain 1 and 15 messages, respectively. None of the postings were discarded, only regrouped.

5.3.2 The VOLUNTEER Corpus

For this adversarial stylometry case study, we compiled a group of 24 adults of minimum 25 years old and as many adolescents between the age of 12 and 14 and invited these volunteers to participate in a live online chat session of 20 minutes. To avoid that their language use would be influenced by the knowledge that their posts would be analysed by researchers afterwards, initially, we did not (completely) inform the participants about the actual objectives of this study. More specifically, the adolescents were asked to test a new chat room by chatting one-on-one with someone from outside their school. No further specifications about their chatting partner were provided. The adults, who were based in a different location, were given the information that they were participating in a study that investigated how reticent the children were with personal information on the Internet and, additionally, whether children could identify an adult posing as one of their peers online. Hence, the adults were given the assignment to pretend that they were

a boy or girl between the age of 12 and 15 and, as such, to try to acquire personal information about their adolescent chatting partner (e.g., name, email address, name of their school, home address and phone number).

For the purpose of this study, we selected adults who are to some extent in touch with the adolescents' culture (e.g., secondary school teachers and educators) or who are trained to question people and steer the conversation in a certain direction (e.g., police officers and psychologists), because this provides more difficult test data for our system. More detailed information about the adult participants and the number of messages they produced during the conversation can be found in Table 5.2. After the chat sessions, we informed all participants about the true nature of this study and the results were made available for cross-curricular school activities.

Table 5.1: Distribution of Netlog users according to age group, gender and region in the NETLOG_SUBSET3.

Age Group	**West-Flanders**		**East-Flanders**		**Brabant**		**Limburg**		**Total per Gender Group**		**Total per Age Group**
	Male	*Female*	*Male*	*Female*	*Male*	*Female*	*Male*	*Female*	*Male*	*Female*	
Min16	390	1,503	640	2,015	1,089	4,018	198	803	2,317	8,339	10,656
Plus20	322	229	570	303	768	360	276	120	1,936	1,012	2,948
Total per Category	712	1,732	1210	2,318	1857	4,378	474	923	4,253	9,351	13,604
	2444		3528		6235		1397				

Table 5.2: Demographics of the adult participants in the VOLUNTEER corpus, together with the number of tokens and messages they produced during their session.

	Age	Gender	Occupation	# Tokens	# Messages
P1	40's	M	Police officer	223	26
P2	40's	M	Secondary school teacher	79	23
P3	20's	M	Student	144	23
P4	40's	F	Senior research associate (linguist)	223	41
P5	40's	M	Secondary school teacher	27	11
P6	30's	M	Senior research associate (linguist)	90	26
P7	30's	F	Educator	167	20
P8	30's	M	Educator	160	44
P9	20's	F	Educator	152	36
P10	20's	M	Secondary school teacher	55	14
P11	30's	F	Secondary school teacher	132	31
P12	20's	M	Secondary school teacher	52	17
P13	30's	F	Secondary school teacher	66	13
P14	30's	F	Secondary school teacher	126	27
P15	20's	M	Computer scientist	134	32
P16	40's	M	Secondary school teacher	84	13
P17	30's	M	Computer scientist	418	77
P18	40's	M	Educator	168	38
P19	40's	F	Secondary school teacher	142	16
P20	60's	M	Retired sports coach	186	41
P21	30's	F	Psychologist	159	42
P22	20's	M	PhD researcher (linguist)	224	48
P23	30's	F	Secondary school teacher	111	24
P24	30's	F	Police officer	167	30

5.4 Message-level vs. User-level Experiments

To enable a valid comparison between the aggregated message-level and the user-level approach, the POST_LEVEL dataset was randomly divided into a 60% training set, a 20% aggregation set and a 20% test set. The USER_LEVEL dataset was split in a 80% training and a 20% test set.

Both test sets contained the same users and Netlog messages. During the splitting, the messages were clustered so that no user was present in two different sets. Distributing users rather than messages ensured that no user in training also appeared in the aggregation or test sets, which prevented overfitting of user-specific features.

Evaluation of the classification system proceeded as follows:

1. The best performing experimental design for age prediction established in the previous chapter was trained on the full training set.

2. To aggregate the predictions on the message level to the user level, the age classifier produced probabilities and labels for the aggregation set. The aggregation method was developed using (*i*) a simple-voting system and (*ii*) an ensemble classifier on these probabilities, and performance was established through five-fold cross-validation on the aggregation set.

3. For final validation of the message-level approach, individual classifiers were trained on the training set, produced probabilities and labels for the aggregation set and the test set; the simple-voting and ensemble models were trained on the probabilities given for the aggregation set, and its predictions taken for the test set.

4. User-level classifiers were trained and evaluated through ten-fold cross-validation within the USER_LEVEL training set. The best performing classifier was then used to produce predictions for the test set, which contained data identical to the test set of the message-level experiments.

5.4.1 The Message-based Approach

The goal of the message-based approach is to obtain the following elements that will later be used to identify false user profiles:

1. A set of prediction models $\mathscr{P} = \{P_1, \ldots, P_i\}$ that output the probability

$$\theta_i(\phi_1, \ldots, \phi_n) = P_i[X = \text{PLUS20} \mid (\phi_1, \ldots, \phi_n)]$$

of each user X belonging to the PLUS20 category given a feature vector $(\phi_1, \ldots, \phi_n)$ obtained from i different messages produced by X.

2. An aggregation strategy $f(\mathscr{P}) = \sum w_i \cdot P_i$ that combines all individual predictions in $\mathscr{P}$. In this strategy, each individual classifier P_i is weighted by w_i according to its performance on the validation set. The system then outputs a final decision on the user level based on a vote so that

$$f = \begin{cases} \text{PLUS20} & \text{if } f(\mathscr{P}) < \tau \\ \text{MIN16} & \text{otherwise,} \end{cases}$$

where τ is experimentally determined.

This study compares two different aggregation strategies: (*i*) a *simple-voting* model, which applies a threshold τ for each Pi (e.g., $\tau = 0.5$) and (*ii*) an *ensemble* model, which takes a weighted vote of all available predictions.

As demonstrated in Chapter 4 (Section 4.5.3), we achieved the best result for age prediction based on only a single message per user by creating a complex model of character 4-grams and combining them with sociolinguistic features, using a Multinomial Naïve Bayes classifier and training on an imbalanced dataset in which each user was labelled as MIN16, PLUS20_FEMALE or PLUS20_MALE and reducing the complex classes to PLUS20 in post-processing. Therefore, we used this experimental design to produce probability outputs for each message in the development and test set of the POST_LEVEL data. Next, as described above, for each user, we combined 15 probability outputs into a single feature vector and calibrated both voting models on the development data.

With regard to the simple-voting strategy, we examined two different approaches: (*i*) taking the average of the 15 probability outputs for the PLUS20 class as the final adult probability for each user (*Thres1*); and (*ii*) calculating the ratio of adult postings to the total number of messages for each user (*Thres2*). As is displayed in Figure 5.1, when using a simple-voting approach, the best F-score of 85.5% for the PLUS20 class was achieved when applying the second strategy with a classification error of 9.4% +/- 0.5%.

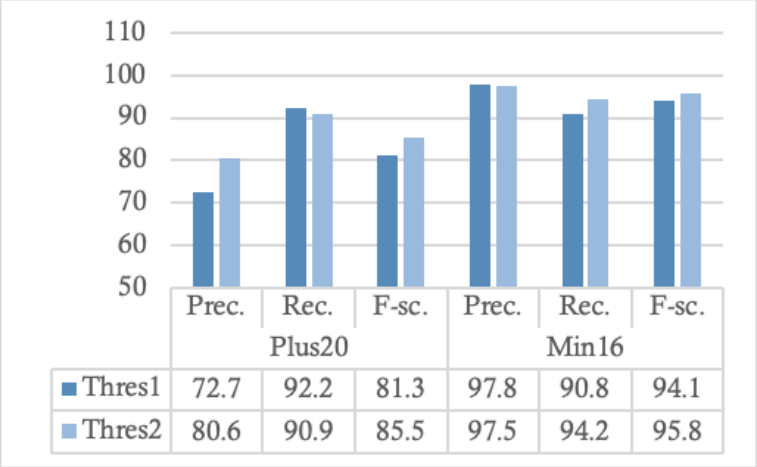

Figure 5.1: Results on the user level after applying different thresholds on the post level predictions of the POST_LEVEL test set.

For the ensemble model, again, we examined the performance of six different machine learning approaches (C-SVC, SGD, NB, *k*-NN, RF and NN). Contrary to the results of the user profiling experiments described in the previous chapter, the Multi-layer Perceptron classifier that was trained using Backpropagation (NN) outperformed all other learners and produced higher results than the simple-voting model described above. This strategy resulted in a precision of 91.1%, a recall of 88.5% and an F-score of 89.8% for the PLUS20 category. This model showed a classification error of 7.6% +/- 0.4%. As demonstrated in Table 5.3, all other ML algorithms produced slightly lower results. In the next section, these results are compared to those of a user-based approach that renders predictions directly on the user level.

Table 5.3: Precision, Recall and F-score for the ensemble model on the test set of the POST_LEVEL dataset.

ML	PLUS20			MIN16		
	Prec.	*Rec.*	*F-sc.*	*Prec.*	*Rec.*	*F-sc.*
C-SVC	90.7	88.0	89.3	94.1	95.5	94.8
SGD	90.7	88.0	89.3	94.1	95.5	94.8
***k*-NN**	90.0	86.6	88.3	93.4	95.2	94.3
RF	89.6	87.9	88.7	94.0	94.9	94.4
NN	**91.1**	**88.5**	**89.8**	**94.3**	**95.7**	**95.0**
NB	90.7	88.0	89.3	94.1	95.5	94.8

5.4.2 The User-based Approach

For each experiment described in this section, we collected, preprocessed and represented all messages from the same user in a single instance vector. This way, the user-based system directly labels users as either an adult or an adolescent and no further aggregation steps are required. To enable a valid comparison with the results of the message-level experiments, we evaluated the model on exactly the same test data as we used in the aggregated message-level experiments.

To allow for a useful comparison between the user-level and message-level experiments, we performed a series of 288 experiments to determine the best performing experimental design when including slightly larger text samples in each document representation (see Chapter 4, Section 4.4). In each experiment, we created a complex model of character 4-gram features and combined them with sociolinguistic features. As can be seen in Table 5.4, this time, the best results for the adult class were achieved by the SGD classifier when training on the binary represented 50,000 most informative features according to the Mutual Information metric, resulting in a 90.0% precision, an 87.6% recall and an 88.8% F-score. This model showed a classification error of 6.8% +/- 0.4%. The other SVM implementation scored slightly lower (88.1% F-score), while the other learners (NB, RF, k-NN and NN) yielded a considerably lower performance.

Table 5.4: Results (%) for the best performing combinations of feature selection and representation methods per machine algorithm for age group identification in the USER_LEVEL dataset.

			Scores					
ML	**Feat. Rep.**	**Feat. Sel.**	PLUS20			MIN16		
			Prec.	*Rec.*	*F-sc.*	*Prec.*	*Rec.*	*F-sc.*
C-SVC	Bin.	MI/χ_2(50K)	90.0	**87.6**	**88.8**	**97.2**	98.2	**97.7**
SGD	Bin.	All feats./χ_2(50K)	91.9	84.9	88.2	96.6	98.7	**97.7**
NB	Bin.	MI (50K)/All feats.	85.3	87.6	86.5	96.3	98.0	97.2
k-NN	Tf-idf	All feats/χ_2(50K)	72.3	77.4	74.7	93.4	96.8	95.1
RF	Abs.	MI(10K)	**94.0**	61.7	74.5	92.5	**99.2**	95.7
NN	Abs./Bin.	MI(50K)/All feats.	86.9	84.5	85.7	93.3	98.6	95.9

5.5 Adversarial Stylometry: Detecting Potentially Suspicious User Profiles

Based on the results of the experiments described in the previous section, both the three-way aggregated message-level (with ensemble approach, see Section 5.4.1) and the three-way user-level classifier (see Section 5.4.2) were trained on, respectively, the entire POST_LEVEL and USER_LEVEL datasets. Subsequently, we evaluated each model on the VOLUNTEER Corpus. The results are shown in Table 5.5.

Remarkably, the results showed a steep decrease in the performance of the user-level classifier: out of 24 imitation passages, it was only able to reveal 14 users as being adults. The remaining 10 adults were wrongly labelled as MIN16, which only resulted in an F-score of 55.0% for the adult age group. Moreover, only 11 children were accurately classified as MIN16. As a result, the aggregated message-level classifier significantly outperformed the user-level classifier when detecting adult volunteers that were asked to pose as adolescents: the post-based age classifier was able to detect 236 out of 706 adult posts, which — after applying the ensemble method — resulted in all 24 volunteers being correctly categorised as PLUS20. Yet, the aggregated post-level approach also rendered 15 false positives, which resulted in a precision of 72.7%, a recall of 100% and an F-score of 84.2% for the adult class.

Table 5.5: Results for the adversarial stylometry experiments on the VOLUNTEER Corpus.

Approach	**Gold Label**	**Plus20**	**Min16**
Message-level	**Min16**	9	15
	Plus20	24	0
User-level	**Min16**	13	11
	Plus20	14	10

5.6 Conclusions and Discussion

At the end of the chat sessions, we invited the adolescent participants to provide some feedback on their experience. Remarkably, when they were asked to guess whether their chatting partner was younger than 16, between 16 and 25, or older, all of them responded "younger than 16". Moreover, 17 children gave away the name of their school, 9 provided their name and surname, 3 gave up their personal email address, 2 agreed to continue the conversation later on Facebook, 1 girl posted her cell phone number and, shockingly, no less than 3 children agreed to meet up at a nearby shopping centre with their unknown chatting partner. These results support the conclusions by [130] stating that children are often careless when communicating with strangers online, especially when they are under the impression that they are talking to one of their peers. Moreover, although at the beginning of the chat sessions we explicitly mentioned that their chat partner was completely unknown to us, none of these adolescents doubted that (s)he was a teenager. These findings help to illustrate that automated methods for predicting age and gender in social media are becoming indispensable to create a safer Internet for children and adolescents.

With regard to the present study, prior research in adversarial stylometry still shows a number of limitations: (a) some studies were performed on computationally modified data instead of on deceptive texts written by humans; (b) the studies that did include obfuscation or imitation passages that were created by human beings were all performed within the framework of authorship attribution studies and did not investigate the feasibility of people imitating the writing style of another sociological group; and finally, (c) all studies were performed on a minimum of 500 words per author and only included formal text genres. Therefore, this study investigated whether a human being is able to modify his or her writing style in a way that it reduces the performance of an age prediction system in the context of social media communications. More specifically, in this chapter, the following research questions were examined:

Q7 Is it feasible to design a system which is able to identify a user's age group even if (s)he imitates the writing style of a different age or gender group?

Q8 Which experimental design leads to a more robust performance when detecting adversarial text passages?

To address these research objectives, we set up a total of 24 one-on-one chat sessions between adolescent and adult volunteers that were asked to pose as adolescents. Additionally, we examined two different approaches to predicting age on the user level: a new post-based approach, in which the classification output on the message level (Multinomial Naïve Bayes) were aggregated to the user level by an ensemble classification model (Neural Network), and a more traditional user-based approach, in which predictions were rendered directly on the user level (linear SVM with Stochastic Gradient Descent). The results showed that — in line with prior adversarial stylometry studies on authorship attribution (e.g.[31]) — when the best performing user-based model was applied to the adversarial passages in the VOLUNTEER corpus, the performance decreased significantly for the PLUS20 age group from a 88.8% F-score on the USER_LEVEL test set to only 55.0% on the adversarial dataset. However, while the aggregated post-level approach only performed slightly better on the non-adversarial POST_LEVEL data than the user-level model, it was able to detect the adult "writing print" flowing through in the adult volunteers' postings, which resulted in all adult volunteers being correctly labelled as older than 20 years old and a final F-score of 72.7% for the PLUS20 age group. This is a relatively small drop compared to the 89.8% F-score that was achieved on non-adversarial data produced on a different online social media platform. These findings seem to confirm the hypothesis formulated in this chapter, i.e., that adults — especially when they were in touch with children's culture, such as teachers — are able to adopt a younger chatting style that can deceive adolescents, but that their innate adult writing style nevertheless flows through in some (parts) of their messages.

Concluding, the results of the adversarial stylometry study presented in this chapter suggest that writing style is to a great extent conscious and can be imitated by other human beings to the point that small clues revealing the author's actual writing style remain undetected by automated systems that are trained on relatively large samples of text. However, as demonstrated in this chapter, these small clues can be detected when training on much smaller text samples, which seems to suggest that writing style is also partially unconscious and which supports the human stylome hypothesis (see Chapter 1, Section 1.1.1.2). In the next chapter, we evaluate the approach on actual child sex offender communications.

CHAPTER 6

IDENTIFYING OFFENDER BEHAVIOUR ONLINE: A CASE STUDY

In the previous chapter, it is demonstrated that it is feasible to automatically detect false adolescent profiles created by adult volunteers. This chapter evaluates the age prediction system on a dataset of actual child sex offender communications and examines if a text mining approach can be used to identify additional aspects of online offender behaviour, such as "grooming" and the sharing of child sexual abuse media.

6.1 Introduction

The proliferation of the Internet has transformed child sexual abuse into a crime without geographical boundaries. While there is scientific debate [142] on whether the online predator is a new type of child sex offender [171] or if those with a predisposition to offend are responding to the opportunities afforded by the new forms of social media [44], empirical evidence points to the problem of Internet based paedophilia as endemic (see also Section 6.2.1).

Recent work shows that nearly half of the offenders who had committed one or more *contact offences*, i.e., they had directly and physically abused children, had displayed so-called "grooming behaviour" [224]. Grooming refers to the process by which an offender prepares a victim for sexual abusive behaviour:

[Grooming is] a process by which a person prepares a child, significant others, and the environment for the abuse of this child. Specific goals include gaining access to the child, gaining the child's compliance, and maintaining the child's secrecy to avoid disclosure. This process serves to strengthen the offender's abusive pattern, as it may be used as a means of justifying or denying their actions. [48]

However, when investigating the possibility to develop automated methods to detect grooming behaviour online, researchers are confronted with a number of issues. First, there is only one benchmark dataset available that contains (English) chatspeak conversations written by child sex offenders: the PAN 2012 Sexual Predator Identification dataset (see Section 6.3.0.1). Yet, because the victims were actually adult volunteers posing as children, it is likely that these conversations are not entirely representative for online predator-victim communications [164]. Secondly, because the seduction stage often shows similar characteristics with adults' or teenagers' flirting, initial studies trying to detect predatory behaviour directly on the user level typically resulted in numerous false positives when they were applied to non-predatory sexually oriented chatspeak conversations in the PAN 2012 dataset [92]. In this chapter, we present a novel approach that combines the advantages of the user profiling system introduced in the previous chapters with text clustering techniques and unsupervised learning to identify different stages of predator grooming.

Secondly, the increasing amount of child sexual abuse media (CSAM) being shared across borders and with apparent impunity leads to new children being found online every day. Each of these children, often from within the family circle of the offender, is a victim of child sexual abuse [96]. The severity of the problem has already resulted in a number of solutions that can monitor such activity. The Child Protection System (CPS) [47] and RoundUp [127, 128], for example, are able to capture data about child sex offender activity and identify CSAM across different peer-to-peer protocols. However, these tools rely on matching the files shared on a network against a hash-value database of known CSAM[1]. As a result, they retrieve thousands of files that have been circulating for several months or even years, but they are not able to identify new CSAM when they are being released on to a network. Nor are they able to detect

[1]Such databases are built through post-hoc forensic analysis of seized computers of offenders.

child sexual abuse media that are not on record. However, one of the main priorities for law enforcement is to identify cases where an offender is actively engaged in the production of *new* CSAM — they can be indicators of recent or on-going child abuse. However, detecting new (or previously unknown) child sexual abuse media is highly challenging, because distributors of CSAM tend to *obfuscate* the illegal content of their shared files. More specifically, they utilise a specialised vocabulary, which contains a variety of keywords, abbreviations, acronyms and even combinations of different languages to avoid (automatic) detection of such files, while making them widely searchable for other offenders. Moreover, this vocabulary proved to be dynamic, i.e. it evolves as existing keywords come to the attention of law enforcement [124]. Finally, detecting new or previously unknown CSAM requires (semi-)automatic analysis of image and video content. However, downloading of all candidate files for automatic image and video analysis is clearly infeasible in, for example, a P2P scenario. Hence, such an approach also requires an intermediate step to reduce the number of candidate files to be downloaded. Therefore, in the second part of this chapter, we describe an intelligent solution to this challenge that adopts advanced text mining techniques to determine the likelihood that a candidate file contains child sexual abuse content based on linguistic information in its filename.

The research presented in this chapter demonstrates that a text mining approach can be used to focus criminal investigations pertaining to child sexual abuse and reduce the amount of time spent looking for digital evidence. Furthermore, an evaluation on verified offender data demonstrates the efficiency of both approaches. The filename categorisation component has been implemented into the iCOP toolkit[2], a software package that is designed to perform live analysis on a P2P network environment.

The rest of this chapter is structured as follows. The next sections discuss background material and related work. Section 6.4 describes the grooming component. The filename classifier is discussed in 6.5 and its integration in the iCOP toolkit is explained in Section 6.6. Finally, Section 6.7 concludes this chapter.

[2] `http://www.research.lancs.ac.uk/portal/en/upmprojects/fp7-icop(e8b3a58b-61d0-42dc-b29c-cfc65c1f37d4).html`

6.2 Background

6.2.1 Trends and Forms of Online Child Exploitation

6.2.1.1 Statistics

Reporting accurate statistics is very challenging in the area of online child exploitation for many reasons. First, the results can only be based on the number of cases that are actually reported to law enforcement. Secondly, because of the novelty of the offence, different definitions of online child exploitation have been adopted across the EU, which makes it difficult to obtain figures on an international scale [96]. Nonetheless, a recent study requested by the European Committee on Civil Liberties, Justice and Home Affairs (LIBE) estimated that for contact sexual abuse[3] the prevalence rates in Europe for girls range from 10% (UK) to 39.8% (Switzerland). For boys the rates were between 6% (UK) and 16.2% (Ireland) [96]. Additionally, according to INHOPE's 2016 report [93], 83% of the victims found in online child sexual abuse media that year were girls, with 40% pre-pubescent and 10% of an infant age.

6.2.1.2 Grooming

Recently, within the framework of the European Online Grooming Project[4], a hypothetical model of online grooming was developed based on the analyses of convicted offender case-files and interviews with experts across Europe. Within this model, three types of child sex offenders were identified according to the level of danger they pose to their victims: the *distorted attachment groomer*, the *adaptable online groomer* and the *hyper-sexualised groomer*. The first type believes that the contact with the victim(s) can be seen as a relationship, the second mainly acts upon own needs and regards the victim(s) as mature and capable, while the third type completely dehumanises the victims to objects and tends to have large collections of images and videos displaying child sexual abuse material and often also interrelates with other online predators [74, 217]. Additionally, prior work by the authors of [74, 96, 123] showed that skilled groomers tend to adjust their grooming methods to fit the targeted child — ranging from (combinations of)

[3]This entails direct, physical sexual abuse.

[4]`http://natcen.ac.uk/our-research/research/european-online-grooming-project/`

Table 6.1: Characteristics of three groups of online grooming offenders (taken from [74]).

Dimensions	Distorted Attachment Groomer	Adaptable Groomer	Hyper-sexualised Groomer
Previous convictions	No	No	Yes
Use of identity	Own	Other	Other
Indecent image use	No	No	Yes
Contact with other offenders	No	No	Yes
Offence-supportive belief	Friendship and love	Exchange compliance	Dehumanised as object
Speed of contact	Long before meeting	Tailored escalation	Fast sex talk and action
Contact method	Personalised contact by phone	Contingent contact approach	Non-personal contact approach
Contact maintenance	Persistence of caring and love	Offers of help and services	Threats of punishment
Offence outcome	All want to meet offline	Some want to meet offline	Some want to meet offline

attention, compliments, affection, kindness, recognition, (digital) gifts and even alcohol, drugs or money to *sextortion*, i.e., threatening to disseminate existing images of the victim[5] — to lower their victims' inhibitions and gain their "consent". Table 6.1 summarises the characteristics of each of these types.

With regard to the content of online communications between offenders and their victims, there is a consensus among experts that the grooming process consists of multiple, recurrent stages, such as gathering information about the victim interests and vulnerabilities, gaining access to the victim, lowering the victim's inhibitions, isolating the victim from adult supervision, initiating the abuse and (potentially) attempting to meet offline with the victim. How long such offenders are able to avoid disclosure is usually determined by how "well" they choose their victims, how proficient they are at identifying and filling their victims' needs, how much time they can invest in the grooming process, whether they manage to seduce and control their victims and, finally, how proficient observers are at recognising and responding to this process [61, 123].

6.2.1.3 File-sharing

The scale of child sexual abuse media trafficking in P2P networks has been investigated by a number of studies involving a timespan of several days [89, 90], weeks [124], months [76, 138, 170, 196] or even years [23, 91]. All these studies showed that, worldwide, hundreds of searches

[5]Such images are often obtained through hacking of the victim's personal computer or through unwanted dissemination of coercively generated images by a victim's peers [96].

for child abuse images occur each second, resulting in hundreds of thousands of CSAM files being shared each year. Moreover, research by [28, 29, 190] based on polygraph results and offenders' self-reports found that 85%, 53% and 55% (respectively) of offenders charged with possession of CSAM had committed one or more contact offences.

Additionally, advertisement websites, such as Backpage and Craigslist, that also offer an abundance of ads for adult services, have enabled an explosion of child sex trafficking. The key tactic used by offenders here is to establish contact with other predators by posting ads that contain legal child images (e.g., of children in swimsuits), while expressing their illegal interests in the attached comments. As a result, selling and trading children and child sexual abuse media online has become a saleable business for different types of criminals, because it is highly profitable and very difficult to prosecute [182].

6.3 Related Work

Although a range of awareness campaigns have already been organised internationally (e.g., the EU Safer Internet Programme[6] and Insafe[7]), only few resources have been employed to investigate novel automated methods to support law enforcement agencies or social network moderators when trying to identify online child sex offenders (see Section 6.3). Additionally, due to both the illicit nature of this topic and the privacy issues that are involved, there is only one website that displays predator-victim chat room conversations: the Perverted Justice website (PJ), which contains over 500 English chat conversations between adult volunteers pretending to be adolescents and as such were approached by an alleged child sex offender. However, for machine learning algorithms to be effective in identifying online sexual predators, they need to be trained with both illegal conversations between offenders and their victims and sexually oriented conversations between consenting adults [164]. Since such data are rarely made public, initial studies [135, 164] only experimented with the PJ dataset. The k-NN classification experiments based on word token n-grams performed by [164] achieved up to 93.4% F-score when identifying the predators from the pseudo-victims. The authors of [172] were the first to

[6] `http://ec.europa.eu/digital-agenda/self-regulation-better-internet-kids` (last accessed on 11/11/13).

[7] `http://www.saferinternet.org/` (last accessed on 11/11/13).

include additional corpora in the non-predator class: they included 85 conversations containing adult descriptions of sexual fantasies and 107 general non-offensive chat logs from websites like `http://www.fugly.com` and `http://chatdump.com`. When distinguishing between 200 PJ conversations and these additional chat logs, the Naïve Bayes classifier outperformed the Decision Tree and the Regression classifier, which resulted in an F-score of 91.7% for the PJ class. The authors of [25] used a corpus of cybersex chat logs and the Naval Postgraduate School (NPS) chat corpus and experimented with new feature types such as emotional markers, emoticons and imperative sentences and computed sex-related lexical chains to automatically detect offenders directly in the PJ dataset. Their Naïve Bayes classifier yielded an accuracy of 92% for PJ predators vs. NPS and 94% for PJ predators vs. cybersex based on their high-level features. However, both [172] and [25] did not filter out any cues that were typical of the social media platforms from which the additional corpora were extracted, which could entail that their models were (to some degree) trained on detecting these cues rather than the grooming content. Moreover, because the high-level features described by [25] were (partially) derived from the PJ dataset itself, these experiments may have resulted in overestimated accuracy when detecting predators from the same dataset.

Recently, the detection of Internet child sex offenders has been extensively investigated in the framework of the PAN 2012 competition, during which efforts have been made to pair the PJ data with a whole range of non-predatory data, including cybersex conversations between adults [92]. Because the PAN 2012 benchmark dataset was heavily skewed towards the non-predatory class, most participants applied a two-stage classification framework in which they combined information on the conversation level to the user level (e.g., [83, 144, 163, 213]). Moreover, apart from one submission that used character n-gram features [169], all other studies used (combinations of) lexical (e.g., token unigrams) and "behavioural" features (e.g., the frequency of turn-taking or the number of questions asked). The best results were achieved by [213], who used a Neural Network classifier combined with a binary weighting scheme in a two-stage approach to first identify the suspicious conversations and, secondly, to distinguish between the predator and the victim. Their system achieved an F-score ($\beta = 1$) of 87.3%. However, during their study, they assumed that "predators usually apply the same course of conduct pattern

when they are approaching a child" [213], which is in contrast with research by [74], which resulted in three different types of predators and, hence, of grooming approaches. Moreover, the PAN2012 dataset was also not cleansed of platform specific cues, which could again have led to overestimated F-scores during the competition. A more detailed overview of the results of the PAN 2012 International Sexual Predator Identification Competition can be found in [92].

6.3.0.1 Grooming Detection

With regard to the content of predatory chat conversations, the authors of [135] were the first to investigate the possibility to automatically detect different stages in the grooming process. Based on an expanded dictionary of terms they applied a rule-based approach, which categorised a post as belonging to the stage of gaining personal information, grooming (which included lowering inhibitions or re-framing and sexual references), approach or none. Their rule-based approach outperformed the machine learning algorithms they tested and reached up to 75.1% accuracy when categorising posts from the PJ dataset into one of these stages. A similar approach was used by [140], whose Naïve Bayes classifier achieved a 96% accuracy when categorising predatory PJ posts as belonging to either the gaining access, the deceptive relationship or the sexual affair grooming stage. The second task of the PAN 2012 competition consisted of detecting the specific posts that were most typical of predatory behaviour from the users that were labelled suspicious during the first task. To this end, most participants either created a dictionary-based filter containing suspicious terms [62, 144, 157, 163] or used their post-level predictions from the predator identification task [83, 105, 111]. The best F-score (β = 1) was achieved by [163], who used a dictionary-based filter highlighting the utterances that referred to one of the following grooming stages: sexual stage, re-framing, approach, requests for data, isolation from adult supervision and age- and child-related references. Their approach resulted in a 35.8% precision, a 26.1% recall and a 30.2% F-score. Finally, [61] proposed a method based on Temporal Concept Analysis using Temporal Relational Semantic Systems, conceptual scaling and nested line diagrams to analyse PJ chat conversations. Their transition diagrams of predatory chat conversations seemed to be useful for measuring the level of threat each offender poses to his victim based on the presence of the different grooming stages.

Although these studies showed promising results, the issue remains that these methods are applied on a corpus that contains conversations between offenders and pseudo-victims. Hence, the adult volunteers that were posing as children could not accede to requests for "cammin", sending pictures, etc. As a result, the PJ dataset contains hardly any conversations by hyper-sexualised groomers, because this type of offender typically does not invest much time in the seduction process and switches to a different victim when his needs are not fulfilled quickly. Moreover, it is highly likely that children would have responded differently to the grooming utterances than the adult volunteers did, which could have influenced the language use of the offenders.

6.3.0.2 Detecting Child Sexual Abuse Media

As detecting known CSAM files is relatively straightforward when a hash-value database is available, initial work in this area mainly focused on the ability to disrupt online child exploitation [97], reliability issues regarding mutable identifiers, such as IP addresses and GUIDs [51, 128, 227] and the identification of key sharers [222]. This has already resulted in a number of tools, such as CPS and RoundUp, that can not only monitor such paedophile activities in P2P networks, but also provide additional features such as geolocation capabilities and centralised databases to assist law enforcement in their international struggle against online child exploitation. Moreover, Internet companies such as Google and Microsoft have created software, such as PhotoDNA [141], that enables law enforcement to detect modified versions of known CSAM[8] (see also [12]). However, none of these tools offers support for identifying new or previously unknown child abuse media.

So far, only few attempts have been made to address this issue. The authors of [59] demonstrated that collaborative filtering techniques that are typically used in recommender systems, can be successfully applied to identify new media in P2P networks of a certain category (e.g., pornography, piracy software and popular music). Their method is based on the assumption that file-sharing traffic tends to cluster around interest, especially when it involves illegal content, such as CSAM files. Hence, they were able to detect previously unknown examples from these categories without analysing their contents or filenames. Secondly, the MAPAP project [124]

[8] A comparative analysis of currently used methods for detecting child abuse media online can be found in [221].

specifically targets peer-to-peer file-sharing networks. There, modelling of user activity and identification of CSAM-related keywords is utilised to identify child abuse media. However, the first system was not tested on verified CSAM data and the latter was not evaluated for the scenario of identifying new or previously unseen CSAM files.

Finally, there are only two studies — to the author's knowledge — that used language analysis techniques to identify CSAM. As mentioned before, the authors of [124] investigated the feasibility to automatically construct lists of CSAM-related keywords. Therefore, in Section 6.5.2 their keyword-based approach is evaluated on a new dataset containing verified CSAM-related filenames. The second study [156] examined whether techniques used for SMS normalisation (see [19]) could also be used to circumvent the issue of language variation or noise in CSAM filenames. Because their work on pornographic versus non-pornographic filename classification showed very promising results, in Section 6.5.2 the filename classification module is also evaluated on this experimental set-up.

6.4 Grooming Detection in Social Media Communications

Both police investigators and social media moderators have a limited amount of time and resources. Hence, they would benefit from a system that presents them with a reduced set of possibly suspicious users, which translates into a high-precision system. On the other hand, it is essential that the system does not miss any potential offenders, and hence, that the recall remains high. So far, prior work has attempted to detect predatory behaviour in social networks directly on the user level (see Section 6.3.0.1), which led to a large amount of false positives (e.g., online flirting conversations between adults) [92].

This section presents an alternative approach to these issues. More specifically, we designed a hierarchic workflow that first classifies each user as either an adult or an adolescent. This first component is composed of the high recall, aggregated post-level age classifier described in Chapter 5, which showed its efficiency when distinguishing between adult and adolescent users, even when the adult is masquerading as a teenager. Secondly, because there are no benchmark datasets available that distinguish grooming from non-grooming messages in online social media

conversations, a grooming detection approach required the ability to describe hidden structures from unlabelled data. Hence, we combined the advantages of text clustering and unsupervised learning methods (see Chapter 2, Section 2.5.2) with a dictionary-based approach to identify different aspects of persuasive language use in online child sex offenders' grooming strategies.

The approach is evaluated on the PREDATOR corpus. This dataset contains Flemish chat room conversations between convicted child sex offenders and their victims, which we manually collected in collaboration with Belgian law enforcement agencies from recently closed court case-files. The rest of this section is structured as follows: the next section provides an overview of the PREDATOR corpus. We present an evaluation of the complete predatory behaviour detection system in Section 6.4.2.

6.4.1 Data

The PREDATOR corpus contains 50 Flemish predator-victim chat room conversations that are written by 13 recently convicted adult male child sex offenders in Belgium. During each police investigation, all suspicious conversations were collected by the local Computer Crime Unit (CCU) through analysing the offender's or the victim's computer and were then added to the case-file to be used as exhibits during trial. For some case-files these data were available on CD-ROMs and could be copied directly from the CCU's analyses. If this was not the case, we copied the printouts of the conversations and applied OCR software[9] to digitalise these data. Additionally, we corrected each utterance manually to avoid errors caused by the digitalisation process. Next, the data was pre-processed as described in Section 4.3.1. For the predator class, this resulted in 9,953 tokens and 2,029 posts. Additionally, we categorised each offender into one of three groomer types (see Table 6.1) according to the available information about the persona(e) the offenders had created during their conversations and the presence (or absence) of threats towards their victim(s). Our analysis showed that 5 out of 13 offenders had posed as an adolescent in at least one conversation and that 3 perpetrators had threatened at least one victim (with, for example, posting their pictures on the Internet) if he or she did not conform to his wishes. For each offender, this information is displayed in Table 6.2, together with the number of conversations, posts and

[9]Wondershare PDF Editor Pro: `https://itunes.apple.com/us/app/pdf-editor-pro/id422542706`.

tokens. We show examples of each groomer type in Table 6.3.

Table 6.2: Number of tokens, postings and conversations per child sex offender in the PREDATOR corpus with additional information about their identity use, the presence of threats towards their victims and their grooming type according to [74].

	Tok.	**Post.**	**Conv.**	**Id.**	**Threat**	**Groomer Type**
Pred_1	366	62	1	own	No	Distorted Attachment
Pred_2	235	45	2	own	No	Distorted Attachment
Pred_3	2.951	360	2	own	No	Distorted Attachment
Pred_4	1.861	396	31	other	Yes	Hyper-Sexualised
Pred_5	211	31	1	other	Yes	Hyper-Sexualised
Pred_6	731	105	4	other	No	Adaptable
Pred_7	780	149	1	other	Yes	Hyper-Sexualised
Pred_8	978	193	7	both	No	Adaptable
Pred_9	910	176	1	other	No	Adaptable
Pred_10	960	235	1	own	No	Distorted Attachment
Pred_11	227	33	1	own	No	Distorted Attachment
Pred_12	146	43	1	own	No	Distorted Attachment
Pred_13	703	138	1	own	No	Distorted Attachment

Table 6.3: Examples from the PREDATOR corpus of each grooming type according to [74] and their English equivalent.

Groomer Type	**Example**	**English**
Distorted Attachment Groomer	*ik heb geprobeert je te vergeten*	*I've been trying to forget you.*
	het lukt niet :-(	*It doesn't work :-(*
	loop je straks eens rond in beeld? :-)	*Will you walk around for the camera later? :-)*
	ik verlang nog steeds naar jou	*I'm still longing for you.*
Adaptable Groomer	*is jou mama nu in de buurt?*	*Is your mom around now?*
	ik zou graag zie of je borste	*I'd like to see whether your breasts*
	groter zijn dan de mijne	*are bigger than mine.*
Hyper-Sexualised Groomer	*leg je cam aan*	*Put on your webcam*
	of ik ben weg	*or I'm gone*
	en je weet wat ik dan ga doen	*and you know what I'll do next*

6.4.2 Experiments and Results

For the first component of the predatory behaviour detection system, we trained the three-way aggregated post-level classifier on the entire balanced POST_LEVEL Netlog subset (see Chapter 5). This age categorisation component was trained on a combination of character 4-grams and the sociolinguistic features introduced in this dissertation (see Chapter 4). When applying the age prediction system on the PREDATOR dataset, all of the 13 offenders — including the offenders who created a false (younger) identity — were labelled as being older than 20.

During pre-processing for the text clustering experiments, we tokenised all offender utterances in the PREDATOR corpus, removed stop words and created document instances utilising bag-of-word features. Tf-idf was used for feature weighting. The utterances produced by Pred_7 (hyper-sexualised) and Pred_9 (adaptable) were extracted from the dataset and used to evaluate the resulting dictionary-based filter described below. We selected these two offenders, because together they produced about 20% of the data available and both of them created a false identity to groom their victims online. All other offender utterances were included in the text clustering experiments.

To investigate how many different grooming stage clusters could be inferred automatically from the PREDATOR data, we applied Ward clustering [215] (as implemented in the Python Scipy package[10]) to enable hierarchical agglomerative clustering and used Cosine distance matrix to perform Ward's linkage.

As can be seen in Figure 6.1, the hierarchical clustering algorithm identified two primary clusters, which are divided into four sub-clusters and four low-level clusters.

Subsequently, we performed K-means clustering using the K-means algorithm that is implemented in the Scikit-learn machine learning package [187]. The algorithm requires a predetermined number of clusters. Therefore, we incorporated the resulting number of clusters (10) from the Ward Clustering analysis in the experiments. Next, to provide a sense of the content of each cluster, we extracted the top-30 n-grams that were nearest to each cluster's centroid. The best results were achieved when we combined uni-, bi- and trigrams and removed all features that appeared in more than 80% of the documents. We show the most relevant features per

[10] https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.cluster.hierarchy.ward.html

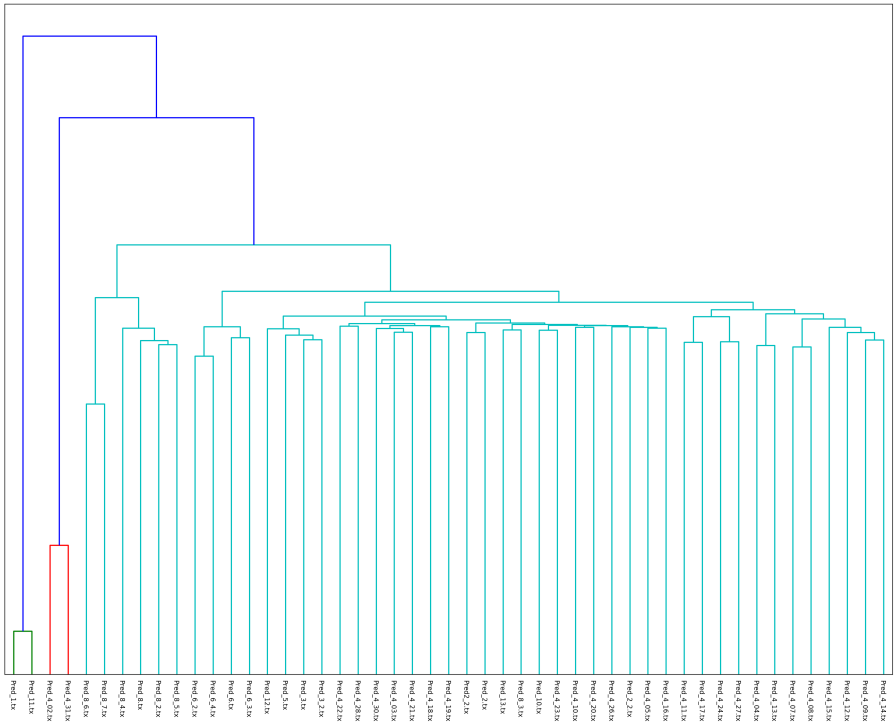

Figure 6.1: Results of the Ward Clustering Analysis performed on the PREDATOR data.

cluster in the K-means clustering analysis in Table 6.4.

Finally, to extract more information about the hidden structure within the PREDATOR data, we used the Latent Dirichlet Allocation (LDA) algorithm (as implemented in the Python Gensim package[11]). We set the algorithm to take 100 passes to ensure convergence and used all data in each pass. Similar to the K-means clustering experiments, the best results were achieved when we included word bi- and trigrams together with the BOW features in the experiments. The results are displayed in Table 6.5.

[11] `https://radimrehurek.com/gensim/models/ldamodel.html`

Table 6.4: 10 Primary Topic Clusters in the PREDATOR data according to K-means Clustering.

Cluster	**Top-30 *n*-grams**
(1)	zien, foto, met_me, spuite, ok, ff, je_met_me, eerst, me_paal, zien, zien_spuiten, jullie_leeftij, wilj, paal, kijken, msn, je_met, een_foto, mss, stuur
(2)	zoet, mooi, aub, oe, prachtig, bh, zo_knap, mmm, zo_mooi, toon, echt, sluit, krijgen, je_prachtig, figuur, je_bent_nog, je_bent, zin, neen, aai_je
(3)	precies, eh, leuke, groter, kun_je, een_half, half, mee, ik_ook, dan_de, verlege, de_buurt, in_de_buurt
(4)	es, zoet, oe, doe, ffe, slipje, snel, zien, wil_je, je_es, laat, en_dan, effe, weg, aai, aai_je, mmm, oe_je, in_je, met_je, ding, mag, oke, op_je, icht, wat_van, ik_ze, cam, goed, ik_wil_je, ik_wil, af, gewoon, van_je, foto's, omhoog, streel_je, echt
(5)	mee, heb_jij_al, geil, wel, jonges, vriend, ies, gezien, ok, je_al, om_je, ga, ni_zo, grote, hoo, heel, ah, dag, al_ies, gevingert
(6)	zien, ok, weet, cam, meer, ben_ik, jah, goed, gewoon, doe, heel, je_cam, oe, ga, zet, toon_je, buikje, wil_je, toon, ik_je, mooi, oke, doet, zei
(7)	inkijk, cup, piepen, allebei, mooi, knap, topje, heeft_die, trui, ook_zo, jammer_da_je, geen, buikje, heeft_de_grootste, wie_heeft_de
(8)	meisje, heb_je, heb_je_een, kiss, geen_dank, geen, lief, een_lief, geeft, go, aardig, heb_je_een, je_een_lief, heerlijk, kus, hoe_zie, heb_je, blond, je_een_foto, blond_blauw, blond
(9)	haar_es, haar_es_achteruit, es_achteruit, soms_aan, bezorg, toon, achteruit, oe_je, je_mij, es_tonen_aan, toon, je_mij, toon, ogen, ga
(10)	lol, mooi, bengel, hoo, oudere, ja_hoo, woon, woon_jij, waar_woon_jij, waar_woon, zal_ik, ver, hallo, ie, da, waar, spuite_al

Table 6.5: 10 Primary Topic Clusters in the PREDATOR data according to LDA Clustering.

Cluster	**Top-30 *n*-grams**
(1)	zoet, es, bent, sluit, krijgen, je_bent, je_bent_zo, bent_zo_kna, komen, :-(_je, janken, neen, trek, zo_kna, kna, toch_niet, bh, schat, waarom, bent_te_mooi, bed_:-)_kom, mooi_voor, ik_lig, schatje
(2)	inkijk, piepen, cup, as, opwinden, allebei, urf_je, urf, mooi, knap, jammer, topje, heeft_die, trui, ook_zo, in_je, es, jammer_da_je, jammer, buikje, geen, die_mooie_cup, wil_wel_es, cht, cht, geen_inkijk, de_grootste, wil_wel_es
(3)	je_cam, cam, weet, neit, cam_aan, zaaag, ik_ben, wel, us, che, h_ik, cam_aan_ja, cam_niet, je_cam_niet, aan_ja, of_ik_ben, of_ik, je_cam_aan, je_cam, meer, doe, asl, dan_ga, niet_aan, ik_weet, niet_meer, het_niet, zien, zag, zie
(4)	echt, als_het, zenn, precies, heel, beetje, foto, zie, nk, mooie, bent_precies, eens_je, je_gezichtje, gezichtje, mag, zien, wil_je, aub, je_bent, graag, je_hebt, helemaal, je_eens, knap, sexy, figuur, kna, mss, ik_ben, toon, of_wat, zwart, omhoog
(5)	meisje, kiss, geeft, heb_je_een, lief, geen, geen_dank, een_lief, kus, heerlijk, dank, een_schat_is, en_geeft, raakt, opgewonden_raakt, opgewonden, geen_dank_heb, dank
(6)	aardig, zacht, mijn_kam, zal_je, kam, geen, mmm, geen_dank, alles_go, heb_je_msn, heb_je_een, je_msn, meissi, lieffi, je_borstjes, liggen, borstjes, over_je, hallo, mij_passen, hallo, mij_passen_lol
(7)	wel, zien, ni, ok, mee, heb_jij, hebt, maar, ma, cam, jah, voor, buikje, ga, toon_je, heb_jij_al, heel, wil_je, toon, gewoon, mag, kijken, an, jij_al, vind, heb_je, es, komen, oe, oke, echt, je_al, filmpjes, gezien
(8)	es, mooi, aub, achtige, je_bent, achtige_ogen, toon, lippen, cht, oe, komen, kom_es_dichterbij, es_dichterbij, dichterbij, zo_mooi, kom_es, komen, gezichtje, je_gezichtje, je_prachtig, figuu, mmm_zo, foto's, je_mij, steeds, nog_steeds, je_bent_nog, bent_nog, zoet, kleedje, prachtig, mmm, ogen, je_es
(9)	es, oe, ale, oe_je, open, slipj, naar_cam, naar, achteruit, effe, haar_es, haar_es_achteruit, haar, soms_aan, open_je, je_best, best, wissen, ale_doe, licht, ffe, handen, cam, op_je, oe_je_haar, je_haar
(10)	geil, hoo, mooi, lol, grote, lol_en, bengel, oudere, ja_hoo, wel, kont, waar_woon, woon_jij, waar_woon_jij, mmm, soms, lekker, alleen_thuis, zal_ik, ie, mm, moe, veel, ver, hallo, sletjes

After performing a qualitative analysis of the results of the unsupervised text clustering experiments, we distinguished the following grooming stages in the PREDATOR data[12]:

(1) **Contact**. Requests for personal information, (such as address, msn, etc.) and isolation from adult supervision (e.g., "alleen thuis?" (*home alone?*));

(2) **Compliments***. Non-sexual and friendly expressions (e.g., "aardig", (*kind*), "mooi" (*beautiful*);

(3) **Approach***. Words and expressions that refer to requests for data (e.g., pictures, using the webcam), imperative forms of verbs that convey the meaning of showing something (e.g., "toon" (*show*)) and expressions that refer to meeting in person;

(4) **Re-framing***. A redefinition of sexually related topics or acts into non-sexual terms (e.g., messing around, practising, teaching, gaining experience) (see also [135]), for example, "al ies gevingert" (*ever masturbated?*, "heb je een vriend" (*got a boyfriend?*), "wie heeft de grootste" (*who has the biggest*);

(5) **Intimacy**. References to more intimate, but not typically sexually related clothing and body parts (e.g., "figuur" (*figure*), "lippen" (*lips*), "buikje" (*belly*));

(6) **Explicit***. References to erogenous parts of the body, mentioning and performing sexual acts and using sexually related adjectives and multi-word expressions (e.g., "geil" (*horny*), "paal" (*pole*/*penis*), "zien spuite" (*see ejaculate*), "borsten" (*breasts*), "slipje" (*panties*), "bh" (*bra*)).

Based on the results of this qualitative analysis, we compiled a dictionary-based filter by (*i*) translating the words from the dictionary by [163] that was based on the analysis of predator utterances in the PJ dataset (see Section 6.3.0.1) into Dutch if they fit into one of these predefined stage categories; by heavily expanding each category with (*ii*) manually selected synonyms and related terms from the Dutch Synonyms website[13]; and, (*iii*) to include typical chatspeak language

[12]The stages marked with (*) were also identified by the authors of [135].
[13]`http://synoniemen.net/`

varieties, with word vectors we produced using the word2vec algorithm[14]. More specifically, we trained the word2vec algorithm on the complete NETLOG, VOLUNTEER and PREDATOR corpora. During this process a vocabulary was constructed from these datasets and vector representations of each n-gram were learnt. In the next phase, we calculated the cosine similarity between a simple mean of the projection weight vectors of each of the 20 most informative n-grams for each grooming stage according to the LDA analysis and the vectors for each n-gram in the trained model[15].

We evaluated the performance of the dictionary-based grooming filter manually on the two remaining chat conversations that were present in the PREDATOR corpus, i.e., the conversations by Pred_7 and Pred_9, which together contained 1,690 tokens and 325 messages. The methodology for detecting the different grooming stages yielded an overall accuracy of 89.8% with a classification error of 10.2% +/- 3.3%. Moreover, the results showed that the approach also performed very well individually for stage (2), (3) and, most importantly for stage (6). The precision, recall and F-score for each stage are shown in Table 6.6.

Table 6.6: Precision, recall and F-score of the grooming filter when detecting the different grooming stages in conversations by Pred_7 and Pred_9.

Grooming Stages	**Precision**	**Recall**	**F-score**
(1) Contact & isolation	50.0	100.0	66.0
(2) Compliments	91.7	84.6	88.0
(3) Approach	84.9	88.2	86.5
(4) Re-framing	94.7	72.0	81.8
(5) Intimate topic	81.8	87.1	84.4
(6) Sexual topic	100.0	90.1	95.2
None	97.6	87.1	92.1

[14]The word2vec implementation in the Python Gensim package: `https://radimrehurek.com/gensim/models/word2vec.html`.

[15]This technique corresponds to the word-analogy and distance methods in the original word2vec implementation [179].

The research presented in this section showed that the age classifier described in Chapter 5 was able to attribute the correct age label to convicted child sex offenders attempting to groom their victims in online social media — even when they created a false user profile. Additionally, the use of unsupervised text clustering techniques contributed to a dictionary-based filter which can be used to identify suspicious conversations between adults and children or adolescents. When combined, these components can enable law enforcement to identify exactly those elements that distinguish predatory grooming behaviour from other online conversations containing similar vocabulary, such as flirting or "sexting" between consenting adults. The next section focuses on a second challenge that was expressed by law enforcement when attempting to identify online child sex offenders [162]: the sharing of new or previously unknown child sexual abuse media.

6.5 Identifying New or Previously Unknown CSAM

The predominant method that investigators use to discover child sexual abuse content in, for example, P2P networks, is a matching of candidate files with known material based on file hashes[16]. Additionally, 70% of their law enforcement experts claimed to use lists of CSAM-related keywords and abbreviations in their investigations[17]. Other information sources, like the image content, were less common.

The remaining part of this chapter demonstrates the feasibility to design an intelligent filtering module that can automatically distinguish between CSAM-related filenames and other P2P material (including adult pornography) while maintaining the complex conditions of a P2P scenario — a large, highly skewed dataset containing extremely short and linguistically noisy text samples. The filename classification module's design is discussed in Section 6.5.1 and evaluated in Section 6.5.2. The architecture of the toolkit itself and its triage model is discussed in Section 6.6.

[16] 96% of [162]'s survey participants claimed to use this method.

[17] In 53% of cases, these were official lists distributed by organisations like InHope, Interpol, IWF, FBI, ICE and CPS, and in the other cases self-created lists.

6.5.1 Approach

Building a filename categorisation module that is sufficiently robust so it can be employed by an automatic environment designed for performing live analysis (such as the iCOP toolkit, see Section 6.6) is a difficult task for a variety of reasons. First, for a machine learning algorithm to be effective in identifying candidate child sexual abuse media based on textual features in their filename, it needs to be trained with both CSAM and non-CSAM filenames. However, there are no CSAM datasets publicly available and crawling for CSAM files directly to acquire training data is illegal. Hence, the research presented in this section was performed in collaboration with specialised law enforcement agencies, which provided access to CSAM-related filenames collected from evidence in closed court cases.

Secondly, the task involved a great number of even shorter text samples (compared to the data described in the previous chapters), which inevitably led to even sparser data. A third challenge lay within the class imbalance inherent to the task: in a P2P environment, for example, the number of non-CSAM files that are being shared highly predominates the number of CSAM files. As most machine learning algorithms are designed to optimise the overall accuracy rate, they have been shown to have difficulty identifying documents of the minority class (see e.g., [189]). Finally, sharers of CSAM tend to create a specialised vocabulary, containing a whole variety of multilingual keywords, and deliberately include linguistic noise, such as abbreviations and acronyms (e.g., “kinderficker”, “kdquality”, “ptsc”) as an adversarial tactic to circumvent detection by law enforcement, while maintaining their availability to other offenders. This poses great difficulties for automated detection techniques — especially because this vocabulary also proved to be dynamic, i.e., it evolves as existing keywords come to the attention of P2P investigators (see [124, 156]). Moreover, supporting multiple languages typically requires sophisticated language identification/translation techniques. This section discusses the text mining approach adopted to address these challenges.

6.5.1.1 Data

Prior research on the Gnutella network [91] reported that 1.6 out of every 1,000 files matched with known child sexual abuse media. Hence, to create a good reflection of reality, for the

filename categorisation module, a highly skewed data distribution was adopted during the learning experiments. More specifically, 10,000 CSAM filenames were matched with 1,000,000 regular filenames from the Gnutella network for the WORLD class and 1,000,000 filenames that were linked to legal pornography media that were taken from *PicHunter*, *PornoHub*, *RedTube* and *Xvideos*[18] for the ADULT class. As mentioned earlier, most filenames are extremely short, containing only 41.6 characters on average (SD = 23.4).

For the classification experiments, the following scenarios that were triggered by practical aspects of law enforcement investigations were designed: (1) detection of CSAM versus regular media (WORLD) and (2) distinguishing CSAM from legal pornographic media (ADULT).

6.5.1.2 Feature types

As mentioned before, distributors of CSAM tend to use multilingual, specialised vocabulary and include spelling variations together with other noise in their filenames to avoid (automatic) detection of their shared files, while making them widely searchable for other offenders. Because the presence of such guarded language use (e.g., "lolita", "childlover", "kdquality", "ptsc") in a filename is highly informative, a dictionary-based filter was created containing a manually extended version of the CSAM-related keyword lists from the MAPAP project [124]. These are referred to as the **CSAM Keyword features**. This filter was further extended with forms of explicit language use (e.g., "handjob"), expressions relating to children (e.g., "kiddie") and family relations (e.g., "daughter") in English, German, Dutch, French, Italian and Japanese. Together, these three categories, i.e., the *explicit language*, the *child references* and the *family references*, form the **Semantic features**. Hence, a filename without any CSAM-related keywords can still become a high-value target with regard to child sexual abuse media when it contains, for example, both explicit language use and references to children (e.g., "handjob11yo"). An example of the feature construction is shown in Table 6.7. The presence of the keyword "pt" (*preteen*) results in a hit for the CSAM keyword features, while "12yo" (*12 years old*) is identified as a reference to a child.

While prior work [124, 156] mainly focused on automatically identifying or normalising typical

[18] `www.pichunter.com`,`www.porno-hub.com`,`www.redtube.com`,`www.xvideos.com`.

keywords that are used by Internet child sex offenders to camouflage their files' illegal content, in this study, a more comprehensive approach is applied by combining the dictionary-based filter described above with other linguistic information. More specifically, all patterns of two, three and four consecutive characters were extracted from the filenames (i.e., **character *n*-gram features**). As can be seen from the example in Table 6.7, this approach allowed for circumventing the issue of alternative keyword spellings: although the actual keyword "lolita" is not present in the example filename, the presence of the "lita" feature could be equally discriminative when training the classifier, because that feature is also present in filenames that do contain the original keyword. Additionally, other potential cues could be picked up by the model, even when they are related to a new or unknown keyword or produced in a language that is not included in our filter.

Table 6.7: Example of a CSAM filename after feature engineering

Original filename	ptl0lita12yo.jpeg
CSAM-rel. keywords	pt CSAM_keyword
Semantic feats.	12yo child_ref
2-gram feats.	pt tl l0 0l li it ta a1 12 2y yo
3-gram feats.	ptl tl0 l0l 0li lit ita ta1 a12 12y 2yo
4-gram feats.	ptl0 tl0l l0li 0lit lita ita1 ta12 a12y 12yo

6.5.2 Experiments and Results

To obtain a reliable estimation of the classifier's performance, a ten-fold cross validation scheme was applied. To enable a comparative analysis between the different scenarios described in Section 6.5.1, we first compiled a complete dataset containing all 2,010,000 filenames and we subsequently created ten training and test partitions. This way, we could vary the training data according to each scenario, but evaluate on the same test data, which still contained all three classes (i.e., CSAM, ADULT and WORLD). Next, for each training partition, we set up four different learning experiments: (*i*) CSAM VS. WORLD, in which we removed the ADULT filenames; (*ii*) CSAM VS. ADULT, where the WORLD data was discarded; (*iii*) CSAM VS. MIXED,

in which we omitted 50% of both non-CSAM classes, and (*iv*) CSAM VS. ADULT VS. WORLD, where the data of the third experiment was reused, but set up as a three-way classification experiment. Hence, the CSAM/non-CSAM ratio (i.e., 10,000:1,000,000) in each experiment was maintained. Additionally, we performed these experiments a second time, balancing the dataset in each training partition but maintaining the original skewed datasets in the test partitions. Finally, to enable a valid comparison to previous work in this area [156], we set up a balanced learning experiment in which only the ADULT and WORLD classes were included.

For classification, we compared the performance of Support Vector Machines (C-SVC and SGD) to Multinomial Naive Bayes (NB), Random Forests (RF), k-Nearest Neighbor (k-NN) and Multi-layer Perceptron (NN). During each learning experiment, parameters were experimentally determined on a development set of each training partition. Because the preliminary experiments showed that a linear kernel was most suitable for dealing with the large, sparse filename dataset, which is in line with [63, 87, 228], we used a linear kernel for training the SVM models.

Contrary to the user profiling experiments described in Chapter 4 and 5, the results in Table 6.8 show that both SVM algorithms significantly outperformed the NB, RF, k-NN and NN classifiers. Because the C-SVC model achieved the best F-score for identifying CSAM-related filenames, this algorithm was used for the remaining learning experiments. To compare the current approach of including additional (noisy) linguistic information to the work of [156] who attempted to normalise the non-standard language varieties in each filename, we set up a balanced learning experiment in which the classifier was trained to automatically distinguish between the WORLD and the ADULT class. Combining character n-gram features with the Semantic features described in Section 6.5.1.2 resulted in a slightly higher accuracy score of 99.3%[19] and a 99.4% precision, a 99.1% recall and a 99.3% F-score for the ADULT class.

However, as expected, identifying CSAM-related filenames proved to be much more challenging. Although when training on the CSAM-related keyword features (see Section 6.5.1.2) the classifier achieved a very high precision score of 93.6%, it also yielded a very low recall and F-score of 6.9% and 12.9%, respectively. In practice, this would mean that out of 10,000 of the verified CSAM-related filenames, 9,310 would remain undetected when using the keyword-based

[19]The authors of [156] reported a best accuracy score of 97.7% for detecting adult pornography filenames.

approach that was suggested by [124]. Therefore, in a next series of experiments we included the Semantic features and character n-grams. Combining all features resulted in a significantly higher recall score of 43.1% and a 55.8% F-score, but also led to a decrease of the precision to 79.9%. As a result, out of 5,175 predicted CSAM-related filenames 932 would be labelled as false positives by law enforcement in the framework of a real-life investigation. The results of these experiments are shown in Table 6.9.

With regard to the different training set-ups, the best precision score was achieved when including all three categories in training. Reducing the ADULT and WORLD categories to NON-CSAM increased the recall to 57.6%, but decreased the precision to 56.9%, which resulted in a slightly higher F-score of 57.2%. This model showed a classification error of 21.6% +/- 0.06%. This decrease in precision could be explained by the fact that two disparate classes (ADULT and WORLD) were combined into a single category. As can be seen in Table 6.10, the two other learning experiments both produced high recall scores, but low precision and F-scores. Balancing the dataset in each training partition, while maintaining a skewed dataset in the test partitions, led to a significantly higher recall score of 80,8%, but the precision and F-score both decreased to 13.7% and 23.7%, respectively.

Table 6.8: Results of the filename classification experiments using different machine learning algorithms.

Scores (%)	CSAM			NON-CSAM		
	Prec.	Rec.	F-sc.	Prec.	Rec.	F-sc.
SGD	52.1	28.1	36.5	99.4	99.8	99.6
C-SVC	**79.7**	43.1	**55.8**	99.5	**99.9**	**99.7**
NB	10.1	**57.5**	17.2	**99.6**	95.6	97.6
RF	8.4	31.1	13.2	99.5	97.5	98.5
k-NN	10.2	57.4	17.4	**99.6**	95.6	97.6
NN	27.9	17.0	21.1	99.4	99.6	99.5

Table 6.9: Results of the filename classification experiments using different feature types.

Scores (%)	CSAM			NON-CSAM		
	Prec.	Rec.	F-sc.	Prec.	Rec.	F-sc.
CSAM-rel. keywords	93.6	6.9	12.9	99.2	99.9	99.6
Semantic feats.	25.6	2.4	4.4	99.2	99.8	99.6
Char. *n*-grams	79.2	41.9	54.9	**99.5**	**99.9**	**99.7**
Combined	**79.7**	**43.1**	**55.8**	**99.5**	**99.9**	**99.7**

Table 6.10: Results of the filename classification experiments using different training set-ups. The set-ups marked with (*) were balanced in training.

Scores (%)	CSAM			NON-CSAM		
	Prec.	Rec.	F-sc.	Prec.	Rec.	F-sc.
CSAM VS. WORLD	2.0	60.1	3.9	99.5	74.8	85.4
CSAM VS. ADULT	18.2	76.6	29.4	99.8	97.0	98.4
CSAM VS. MIXED	56.9	57.5	**57.2**	99.6	99.6	99.6
CSAM VS. ADULT VS. WORLD	**79.7**	43.1	55.8	99.5	**99.9**	**99.7**
*CSAM VS. WORLD	2.6	86.5	5.12	99.8	72.5	84.0
*CSAM VS. ADULT	5.5	**87.5**	10.4	**99.9**	87.0	93.0
*CSAM VS. MIXED	10.0	85.7	17.9	**99.9**	93.3	96.5
*CSAM VS. ADULT VS. WORLD	15.0	80.5	25.3	99.1	94.4	96.7

In this study, we showed that it is feasible to design an intelligent filtering module that can automatically distinguish between CSAM-related filenames and other (adult) P2P material under the complex conditions of a P2P scenario — a large, highly skewed, sparse dataset. Although this approach significantly outperforms the standard keyword-based approach, a false positive rate of 20.3% indicates that a decision from the filename classification module is still insufficient to label a candidate file as CSAM. Hence, a highly precise image classification module is required as a second step in the analysis. The iCOP toolkit [162] combines these two types of media analyses, while disregarding files with known hash values, to flag the most pertinent candidates for new or previously unknown child abuse media. We provide an overview of the toolkit's design in the next

section.

6.6 The iCOP Toolkit

We integrated the filename classification approach described in this chapter into the iCOP toolkit[20], a software package designed to assist law enforcement in identifying and prioritising new or previously unknown child sexual abuse media in P2P networks. As is shown in Figure 6.2, the toolkit has two major components: the P2P Engine and the iCOP Analysis Engine.

The P2P engine provides functionality to monitor public traffic on Gnutella, but other monitors can be plugged into the engine as well. The monitor extracts information such as IP addresses, filenames and hash values of files, together with meta data, such as when a particular peer was last seen sharing a file. The latter is essential to identify the originator of a file after it has been labelled by the toolkit as containing new or previously unknown child sexual abuse content. This information is passed on to the iCOP analysis engine, which undertakes the following steps:

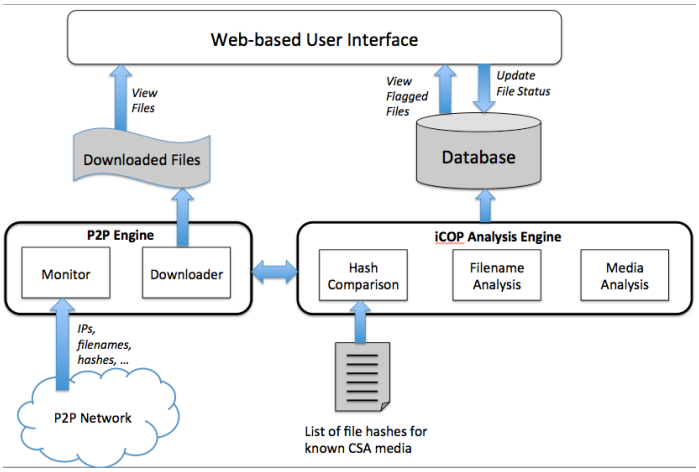

Figure 6.2: Overview of the iCOP toolkit.

[20]The iCOP project was funded by the European Commission Safer Internet Programme project (SI-2010-TP-2601002), *iCOP: Identifying and Catching Originators in Peer-to-Peer Networks*. Research was performed by A. Rashid, C. Fischer and C. Peersman (Lancaster University), C. Schulze (DFKI, Germany) and M. Brennan (Cork University).

1. It compares the hash values of files to a list of known hashes. Such hash value lists are established by law enforcement when child sexual are seized. This filtering mechanism ensures that the system disregards known CSAM. Although the user interface does indicate when a peer is sharing known CSAM, the toolkit does not download or process the files given the focus on identifying new or previously unknown child sexual abuse media. This significantly reduces both storage and computation requirements. Currently a file of SHA1 hashes in base-32 (one hash per line) is used, because this is the most common format in which law enforcement store hash values for CSAM. As a result, the design enables law enforcement officers using the toolkit to plug in their own hash value lists without substantial effort to import them into a specific format or database.

2. The names of files that do not occur in the known hash list are then passed on to the filename classifier (see Section 6.5) for identifying their likelihood of containing CSAM. This is the first step of the automatic CSAM analysis. Filenames that are deemed to be non-CSAM are discarded.

3. Files that are flagged by the filename classifier as potentially containing child sexual abuse content are passed back to the P2P engine for downloading. The downloaded files are then piped back to the iCOP analysis engine and are analysed by the media classifier — the second step of the automatic CSAM analysis — to determine if the content indeed contains child abuse images. When combined with the filename classification module, the system yielded a false positive rate of only 7.9% for images and 4.3% for videos.

The results of the analysis are stored in real-time in the database. An investigator can login to the GUI to access the iCOP "dashboard", which automatically triages the results to flag the most pertinent candidates for CSAM as the highest priority. More specifically, the main table displays details of the connections sharing the greatest number of suspected files based on the results of the image analysis module. The table can also be sorted according to total number of shared files, number of suspicious filenames, or number of shared files known to be CSAM. Additionally, the investigator can view thumbnails as well as the full media files to verify whether the flagged items indeed contain child sexual abuse content. If so, these items can be marked

"confirmed" by the user and are fed back into the hash database so that they are considered to be known CSAM in future searches.

Furthermore, the toolkit GUI is designed around a list of connections, which maps closely to the way P2P software works. A connection is defined as:

$$connection = IP\ address + Port + GUID$$

Each connection is assumed to be a single user sharing a given set of files from a specific location. This is in contrast with an IP address alone, which could potentially be shared by multiple users (e.g., several machines in a home) or a GUID alone, which could potentially be used from different locations (e.g., work, home, travel). The toolkit can display files shared by a particular IP or a particular GUID. Hence, an investigator can easily view which connections are related via a common IP address or GUID. As mentioned above, the most pertinent candidates are flagged to the user as high priority via the dashboard. Given legal constraints governing law enforcement, the toolkit's modules (i.e., the filename and the media classifier) can also be integrated separately into extant investigative workflows and configured to focus on particular geolocations (e.g., a particular country or region).

Finally, in order to accommodate the requirements for the development of victim-centric approaches in jurisdictions that are restricted by statutory proscriptions on the enactment of pro-active surveillance strategies (see [162]), the toolkit was developed in accordance with a modular design that permitted flexibility in any operational application of the iCOP toolkit. This configuration enabled domain law enforcement to deploy iCOP as an integrated solution alongside existing P2P tools in proactive monitoring contexts, or to adapt its media or filename classifiers to the identification of new CSAM files during reactive investigative activities (e.g., incorporating iCOP's component classifiers into triage investigation procedures during post hoc forensic examinations of seized hardware).

6.7 Conclusions

Whether charged with enforcing the law in respect of broader offences of possession and distribution, or with the apprehension of producers of child abuse media, the identification of contact sexual abuse and abuse victims were cited as paramount concerns for P2P investigators [162]. This finding resonates with earlier observations that a primary goal of P2P investigations is to catch child abusers and help children that are being sexually victimised, rather than simply detecting and confiscating images in the context of possession offences [127, 128]. However noble, these objectives are difficult, nigh impossible to realise using the current state-of-the-art tools, which offer little support for the identification and prioritisation of high-risk targets — such as those grooming children in online social media or those distributing new or previously unknown child sexual abuse media that may indicate recent or ongoing child abuse. As a result, this chapter examined whether automated text-based methods can be employed to assist law enforcement in their child protection investigations. More specifically, the following research questions were (again) examined:

Q7 Is it feasible to design a system which is able to identify a user's age group even if (s)he imitates the writing style of a different age or gender group?

Q8 Which experimental design leads to a more robust performance when detecting adversarial text passages?

Q9 Can a text mining approach be used with sufficient reliability to identify media which contain child sexual abuse content based on linguistic features in its filenames, despite the even shorter text samples and obfuscation techniques used by offenders to avoid detection by law enforcement?

Q10 Can a text mining approach be used to identify different stages of grooming which are employed by child sex offenders to deceive their victims?

In this chapter, a new set of tools was presented: (1) a predator detection component, which can be used to detect predatory behaviour in online social media, such as the use of false adolescent

profiles and grooming; and (2) a filename classification component which is already integrated in a software package that is designed to highlight sharers of new or previously unknown child sexual abuse media in P2P networks. Both components were evaluated on real offender data, which showed high degrees of accuracy. Moreover, the modular design of each component enables law enforcement agencies to integrate (parts of) the software into their extant investigative workflows or to add new extensions. The tools are currently being made available to law enforcement. Interested parties should contact the author for more information.

CHAPTER 7

Conclusions and Future Research

Although the last decades of research have brought substantial innovation to the field of computational stylometry, it is still dominated by studies performing author profiling on balanced datasets containing a considerable amount of formal text samples for each author (e.g., entire books or essays). As a result, it is uncertain how the proposed approaches will scale towards other text genres and different data distributions.

The increasing amount of child sex offenders exploiting social media and peer-to-peer networks with apparent impunity and the scarcity of research into technological approaches that are able to identify such criminal behaviour online provided a key motivating factor for focusing this work on three aspects that are becoming increasingly important in the field and that are essential for designing applications for online child protection, namely the scalability of a text mining approach when confronted with online social media communications that contain (a) linguistically noisy text samples, (b) highly sparse feature vectors and (c) adversarial (or deceptive) text passages.

The first part of this chapter summarises the main findings of this dissertation with regard to these three aspects. Subsequently, the implications of this study for developing forensic applications are discussed in Section 7.2. Next, future research perspectives are formulated in Section 7.3. Finally, this thesis closes with a few concluding remarks in Section 7.4.

7.1 Research Objectives Revisited

7.1.1 Noisy Data

Chatspeak presents an important challenge for present-day text mining research. Apart from its (theoretical) relevance in, for example, linguistics, it currently impedes a number of practical applications, such as user profiling and opinion and sentiment mining. In this thesis, the NETLOG Corpus was presented, a comprehensive collection of over 1.4 million messages, which was collected for this study together with its users meta-information, in collaboration with the Belgian social networking platform Netlog. Additionally, it incorporates nearly 900,000 postings produced by adolescents without any domain restrictions, making this dataset a unique resource both for sociolinguistic research and for developing child protection applications.

The presence of linguistic noise in social media communications is said to provide significant challenges for text mining research, because many off-the-shelf NLP tools fail to correctly analyse this anomalous input. Hence, most text mining approaches tend to discard all non-standard language varieties or attempt to normalise them to improve feature extraction procedures. However, systematic studies of the level of linguistic noise in Flemish online social network communications and its correlation with demographic features have been a lacuna in the field. Therefore, in this thesis a forward stepwise mixed-effects logistic regression analysis was set up to examine the effects of age and gender on the production of two types of non-standard language use: newly incoming non-standard forms that are inherent to the medium and written representations of older regional and dialect forms that are typical for colloquial speech. These analyses showed that age had a significant non-linear effect on the production of both types, indicating that the Adolescent Peak Principle, which was reported in previous spoken discourse studies, is also reflected in written communications on social media.

Because these results illustrated the potential of such (un)consciously made linguistic choices to contribute to automated methods for user profiling, instead of discarding or normalising them, a novel feature engineering method was developed to automatically extract *sociolinguistic* features, i.e. information on each feature's "standardness" together with paralinguistic and non-verbal features such as character flooding and emoticons. Interestingly, when these features were

combined with character n-gram features, together, they outperformed all other (combinations of) feature types when detecting age and gender on the message level and when identifying false user profiles on the user level. Hence, these results not only demonstrated their reliability when dealing with limited data, they also proved to be useful for tracing stylometric evidence beyond topic and across different online platforms.

7.1.2 Data Size

In the context of short text mining, it is standard practice to reduce data sparseness by either grouping multiple text samples into one document instance or by incorporating additional word level concept information obtained from external sources (e.g., WordNet). However, one of the key considerations for user profiling in social media when conducting cybercrime investigations is that there is no control over the new data that will be presented to the system, which could be limited to a single message.

This thesis provided a systematic study of different aspects of experimental design to assess their scalability towards highly sparse data as can be found in (*i*) online social media communications (12.7 tokens on average per message) and (*ii*) filenames (41.6 characters on average per filename) distributed on P2P networks. The findings described in this dissertation showed that, despite the challenging characteristics of these text samples, it is feasible to improve significantly upon random baseline performance for both automatic user profiling and child sexual abuse media identification.

7.1.3 Adversarial Data

Contemporary computational stylometry research typically focuses on identifying and extracting linguistic features which are potentially discriminative for an author's stylome and developing an efficient computational model which includes these features to automatically determine an author's identity or profile. However, when developing automated methods for user profiling "in the wild", it is essential that the potential for adversaries is taken into consideration. Therefore, three adversarial stylometry experiments were designed in the context of the present study, in which the system was evaluated on a dataset of adults posing as adolescents (imitation), a

dataset containing chat room conversations by recently convicted child sex offenders (obfuscation/imitation) and, finally, a dataset of actual CSAM filenames containing specialised vocabulary to circumvent detection by law enforcement (obfuscation). So far, none of the prior studies in author/user profiling has evaluated their system on such adversarial passages.

In line with previous studies on detecting imitation in authorship attribution, the classification models which were trained directly on the user level demonstrated a significant decrease of the performance when evaluated on the VOLUNTEER Corpus. Therefore, it was our hypothesis that the number of clues revealing the authors' own writing style were too limited to stand out between the majority of imitated features when training on larger text samples. As a result, a novel approach was presented, in which predictions on the message level were aggregated to the user level by an ensemble voting model. Contrary to the traditional user-based approach, the aggregated message-based classifier was able to successfully detect the adult "writing print" flowing through most postings, which resulted in an accurate categorisation of all adults in both the VOLUNTEER and the PREDATOR dataset. These results seem to confirm that writing style is to a great extent conscious and can be imitated by other human beings to the point that small clues revealing the author's actual writing style remain undetected by automated systems that are trained on relatively large samples of text. However, as demonstrated in this thesis, such small clues can be detected on the message level, which seems to suggest that writing style is also partially unconscious and supports the human stylome hypothesis developed by [206].

7.2 Implications for Digital Forensic Applications of Text Mining in Online Social Media

The potential to accurately analyse and reconstruct a digital forensics incident depends on the capability of a system to cost-effectively collect and preserve incident-related data. Although this capability, which is also referred to as *forensic readiness*, plays a crucial role in supporting digital forensic investigations, most text mining studies pay little or no attention to how a forensically ready system can be developed or what its requirements are. This section operationalises forensic readiness goals as described by [158] by using the tools described in this thesis.

The authors of [158] identified the following forensic readiness goals: *Availability, Relevance, Non-Repudiation, Legal Compliance, Completeness, Minimality* and *Linkability*.

Availability. The availability of data is dependent on preservation and accessibility of the data. Moreover, in order to facilitate data retrieval, the information preserved should also include the meta-data. Because some data is ephemeral, such as network traffic, preservation should be performed pro-actively — before any investigation is initiated. Finally, forensic readiness also implies that stored data can be accessed by law enforcement any time this is required. Hence, it is obvious that any trained text categorisation model would require a pipeline in which social media data are first collected and stored, together with meta-data, such as profile information, email address, IP address, etc. For example, the iCOP toolkit[1] includes a P2P engine, which collects data (e.g., filename, file hash, IP address, port, client ID, geolocation, ...) from the Gnutella P2P network and uses a range of different strategies to disambiguate these data[2]. Secondly, at the end of the pipeline should be some type of "intelligent dashboard", which enables a police investigator to triage and prioritise the incoming data efficiently. The objective of pro-active data storage is, however, very difficult to achieve in social media contexts. For example, downloads in P2P networks can stall if computers go off-line or suspects can use a social network platform to establish contact with their victims and change to more private environments to hide their criminal intentions from law enforcement or social network moderators.

Relevance and Minimality. To be useful as evidence, preserved data should be relevant to the incident under investigation and the preserved data should occupy minimum storage space. Both objectives are critical to guarantee a minimum amount of resources being used during an investigation. The tools presented in this thesis demonstrate the potential of a text mining approach for reducing the amount of time and resources that are spent looking for digital evidence. More specifically, the user profiling component was designed to automatically flag deceptive profiles out of thousands of user profiles that are active on any given social network, while the grooming detection component is able to highlight persuasive tactics used by child sex offenders among hundreds of thousands of communications that take place in such environments.

[1]The filename classification module described in Chapter 6 was incorporated in the iCOP toolkit.

[2]For example, the toolkit stores all meta-data in separate tables in order to easily identify which client IDs are being used by the same IP, etc. Additionally, iCOP stores timestamps for all files which would allow an Internet service provider to identify who was assigned the IP address at a given time.

Additionally, the filename classifier, when combined with media analysis techniques, has proven its usability for detecting child sexual abuse media on P2P networks. Although each component was successfully tested on actual offender data, due to the nature of any text mining approach — false positive errors are inevitable — the relevance criterion can never be fully met. For example, webcam videos displaying revealing images of children without any adult interaction might be flagged as CSAM-related, but according to law, it is not. This characteristic emphasises the role of any automated method in a digital forensic context as a decision supporter, not a decision maker.

Linkability. Data should be linkable with other pieces of evidence, such as witness statements. To meet this goal, again, additional methods, aside from the text mining components, are required that can analyse IP addresses, port access, etc. to identify unique connections. However, different text mining components can be used to help achieve this criterion. For example, when a suspicious user is flagged for sharing illegal media on a P2P network, more evidence could be found when automatically analysing his/her social network profiles, or vice versa. With regard to the iCOP toolkit, the software not only displays the number of potentially new/previously unknown CSAM files that are being shared by a particular connection, it also provides information about the number of shared files that are known to contain child sexual abuse content. This way, false positive results of new CSAM-related media, like a revealing video of a child on the beach, can still become informative in a police investigation when the same connection is also sharing previously confirmed CSAM files.

Non-Repudiation. Non-repudiation refers to the legal admissibility of digital evidence, i.e. its ability to be accepted in a court of law. To achieve this goal, preserved data should not be tampered from the time of acquisition until its final deposition. Additionally, the preserved data should provide high assurance about its authenticity and only authorised parties should be able to access it. Finally, the chain of custody should be maintained, meaning that every change in the control, handling, possession, ownership or custody of a potential piece of evidence should be documented. As a result, this goal can only be achieved when a text mining component is integrated into a GUI, which is exclusively accessed by police investigators with the proper login, or by integrating the component into extant investigative workflows.

Legal Compliance. Preserved data should ensure compliance with existing regulations,

which may vary depending on the jurisdictions in which the incident under investigation has occurred. Such regulations can affect the privacy of the data, how long it can/should be retained and how it can be accessed or collected. Hence, contrary to the larger part of contemporary NLP applications (e.g., the LIWC api), a forensic application should have a modular design, permitting flexibility in any operational application and allowing for each component to be used separately if required. For example, to this date, Belgian police investigators are still restricted by statutory proscriptions on the enactment of pro-active surveillance strategies. Hence, the iCOP toolkit, which was designed to perform pro-active analyses on P2P networks, cannot be employed. However, the modular configuration of the toolkit does allow for investigators to adapt its filename and media classification components to the identification of child sexual abuse media during reactive investigative activities (e.g., incorporating the components into triage investigation procedures during traditional digital forensic examinations of seized hardware).

Based on the theoretic framework of forensic readiness formulated by [158], this section provided a discussion on the requirements of a text mining approach to be efficiently used within a legal context, such as digital forensic investigations pertaining to child sexual abuse. This framework allowed for analysing the trade-offs at play, which influence the satisfaction of certain forensic readiness requirements within the tools presented in this thesis and the iCOP toolkit.

7.3 Future Work

This thesis has demonstrated the complexity of detecting adversary behaviour “in the wild”, as well as the challenges of applying such an approach in a digital forensic framework. Hence, ideas for further research were inspired by the limitations of the present study and by the observations made above.

First of all, this dissertation investigated a variety of interacting factors that can affect the performance of a text mining approach and influence its scalability towards a real-life application for (Flemish) Dutch. As a result, future research will include the development of a cross-language approach, in which comparative analyses will be presented for other languages. Hence, it can be investigated which aspects of adversarial language use can contribute to a more general

deception detection model that could also be portable to other languages. To achieve this goal, new datasets will need to be compiled in collaboration with law enforcement agencies and other international organisations, such as the UNODC[3], for the user profiling, grooming and filename detection components.

Secondly, this study focused on developing new models for online child protection. As mentioned before, creating a false identity in online social media is a tactic used by other cybercriminals as well. Therefore, future work will include collecting new datasets and retraining the tools to be useful for e.g., detecting fraudsters on online dating websites. Such an approach will require more fine-grained age categorisation models for the adult class, because adolescents are typically not targeted by such services.

Third, to this date, it is still unclear which characteristics are significant predictors for committing different types of cyber crimes, if cyber offenders show specific patterns in their criminal behaviour and what the major motivating factors are at different stages in their careers. Therefore, our future research will focus on developing new text mining approaches that can be used to uncover such identifying characteristics of cyber offenders, which will not only provide insight into their online behaviours, motivations, police evasion tactics and pathways in and out of cybercrime, but will also provide law enforcement with new tools to identify and track down suspects.

Finally, future work will also include a more formal characterisation of forensically ready systems and to investigate different types of reasoning that can be adopted to assess the satisfaction of forensic readiness goals. Additionally, different aspects related to the implementation of a forensically ready system will be investigated.

7.4 Concluding Remarks

This thesis has highlighted the potential for a systematic approach which automates child protection services in online social media. The tools presented in this dissertation provide a key step in developing automated methods that are able to assist law enforcement investigations pertaining to online child sexual abuse by (*i*) detecting victims at acute risk, (*ii*) assigning

[3] `http://www.unodc.org/`

degrees of importance and urgency to items of evidence in order to assess offenders' potential danger to society and (*iii*) find useful evidence in a timely manner. By tying the examination of methodological issues in the field of text mining and computational stylometry to more practical objectives of developing new techniques to help increase online child protection, this work, hopefully, will contribute to the development of useful applications for supporting cybercrime investigations.

APPENDIX

A.1 Sample of the permission form used for creating the VOLUNTEER Corpus (in Dutch)

The following document was signed by the parents of each adolescent volunteer who participated in the adversarial stylometry study described in Chapter 5.

Beste ouder(s),

In het kader van een wetenschappelijk project van de Universiteit Antwerpen over het gebruik van chatapplicaties bij jongeren zijn wij op zoek naar leerlingen tussen 11 en 13 jaar oud die een nieuw chatprogramma willen uittesten op school. Concreet zullen wij hen vragen om op woensdag 23/10/13 om 12.00u deel te nemen aan een online chatgesprek van ongeveer 20 minuten, waarbij ze de verschillende functies van het programma kunnen uitproberen. Daarna wordt hen nog een korte vragenlijst aangeboden, waarin zij feedback kunnen geven op de vormgeving, gebruiksvriendelijkheid, enz. Om hen te bedanken voor hun deelname, bieden wij hen een gratis filmticket aan. Alle gegevens die binnen deze studie worden verzameld zullen worden geanonimiseerd en nadien uitsluitend worden gebruikt voor wetenschappelijk onderzoek binnen het project.

Indien uw kind mag deelnemen aan ons onderzoek, gelieve dan de onderstaande strook in te vullen, te ondertekenen en **ten laatste op maandag 14/10/13** terug te bezorgen aan het secretariaat van de Middenschool. Voor meer informatie en verdere vragen rond het wetenschappelijk onderzoek kan u steeds terecht bij Claudia Peersman via claudia.peersman@ua.ac.be.

Wij danken u bij voorbaat voor uw medewerking.

Claudia Peersman Prof. Dr. Walter Daelemans

Universiteit Antwerpen lerenisleven.be
www.ua.ac.be

Ik, ondergetekende, ……., ouder van de leerling(e) …… uit klas ……..
geef hierbij de toestemming aan mijn kind om op 23/10/13 om 12.00u deel te nemen aan een online chatgesprek van ongeveer 20 minuten dat kadert binnen een wetenschappelijke studie van de Universiteit Antwerpen en nadien een vragenlijst hierover in te vullen. Bovendien geef ik de toestemming aan de onderzoekers om het chatgesprek en de vragenlijst, na anonimisering, verder te gebruiken voor wetenschappelijk onderzoek. Mijn kind zal als dank voor zijn/haar deelname een gratis filmticket ontvangen.

Datum:

Handtekening ouder(s):

The following document was completed by the adult volunteers who participated in the adversarial stylometry study described in Chapter 5.

Beste deelnemer,

In het kader van ons onderzoeksproject dat gaat over de veiligheid van kinderen en jongeren in chat rooms en op sociale netwerken zoals Facebook en Netlog organiseren wij deze woensdagmiddag (23/10/13) een chatsessie van ongeveer 20 minuten. Concreet is het de bedoeling dat u **om stipt 12u inlogt** op de link die u morgen in de voormiddag zal worden toegestuurd. In deze chat room zal u samengebracht worden met een leerling van het eerste middelbaar. Aan deze leerling werd enkel gevraagd om een nieuwe chatomgeving uit te testen en er nadien feedback over te geven. De opzet van het experiment is echter dat u zich probeert voor te doen als iemand (jongen/meisje) van ongeveer dezelfde leeftijd. U mag dus over alles, inclusief over uw leeftijd en geslacht, liegen. Vervolgens willen we u vragen om tijdens dit gesprek te trachten zoveel mogelijk persoonlijke informatie te weten te komen van deze leerling (naam, woonplaats, naam van de school, e-mail, telefoonnummer, etc.). Wanneer de 20 minuten voorbij zijn, zal u een melding zien verschijnen en mag u de chatsessie afsluiten.
Met dit experiment willen we nagaan hoe gemakkelijk jongeren dergelijke gevoelige informatie prijsgeven aan onbekenden via het internet. Bovendien zullen we nadien testen of de leerlingen er enig besef van hadden dat ze eigenlijk met een volwassene aan het chatten waren.
Door deel te nemen aan het experiment verklaart u zich akkoord dat uw chatgesprek na anonimisering verder mag worden gebruikt binnen het onderzoeksproject. Voor uw deelname ontvangt u twee Kinepolis filmtickets.

Gelieve ook onderstaande gegevens verder aan te vullen en terug te mailen:

Naam:
Voornaam:
Adres:
GSM nr.:
Geboortedatum:
Geslacht:
Beroep:

De tickets zullen na uw deelname worden opgestuurd naar het ingevulde adres.

Wij danken u nogmaals voor uw medewerking.

Hoogachtend,

Claudia Peersman Prof. Dr. Walter Daelemans

Universiteit Antwerpen
lerenisleven.be
www.ua.ac.be

BIBLIOGRAPHY

[1] AFROZ, S., BRENNAN, M., AND GREENSTADT, R.

Detecting hoaxes, frauds, and deception in writing style online.

In *Security and Privacy (SP), 2012 IEEE Symposium on* (2012), IEEE, pp. 461–475.

[2] AGGARWAL, C., AND ZHAI, C.

Mining text data.

Springer Science & Business Media, 2012.

[3] AGGARWAL, C. C., AND ZHAI, C.

A survey of text clustering algorithms.

In *Mining text data*. Springer, 2012, pp. 77–128.

[4] AHA, D.

Lazy learning.

Springer Science & Business Media, 2013.

[5] AL ZAMAL, F., LIU, W., AND RUTHS, D.

Homophily and latent attribute inference: Inferring latent attributes of Twitter users from neighbors.

ICWSM 270 (2012).

[6] AMERICAN PSYCHIATRIC ASSOCIATION.

Diagnostic and statistical manual of mental disorders (DSM-5®).

American Psychiatric Pub, 2013.

[7] ANDROUTSOPOULOS, J., AND ZIEGLER, E.

Exploring language variation on the Internet: Regional speech in a chat community.
In *Language Variation in Europe: Papers from the Second International Conference on Language Variation in Europe, ICLaVE* (2004), vol. 2, pp. 99–111.

[8] ARGAMON, S., KOPPEL, M., PENNEBAKER, J. W., AND SCHLER, J.
Automatically profiling the author of an anonymous text.
Communications of the ACM 52, 2 (2009), 119–123.

[9] AUER, P.
Europe's sociolinguistic unity, or: A typology of European dialect/standard constellations.
Perspectives on variation: Sociolinguistic, historical, comparative (2005), 7–42.

[10] AUER, P., AND HINSKENS, F.
The convergence and divergence of dialects in Europe. New and not so new developments in an old area.
Sociolinguistica 10 (1996), 1–30.

[11] BAKHTIN, M.
The dialogic imagination: Four essays.
University of Texas Press, 2010.

[12] BALUJA, S.
Building software tools to find child victims, 2008.

[13] BAMMAN, D., EISENSTEIN, J., AND SCHNOEBELEN, T.
Gender in Twitter: Styles, stances, and social networks.
CoRR abs/1210.4567 (2012).

[14] BANERJEE, S., RAMANATHAN, K., AND GUPTA, A.
Clustering short texts using Wikipedia.
In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (2007), ACM, pp. 787–788.

[15] BARON, A., RAYSON, P., GREENWOOD, P., WALKERDINE, J., AND RASHID, A.

Children online: A survey of child language and CMC corpora.

International journal of corpus linguistics 17, 4 (2012), 443–481.

[16] BARON, N.

Computer mediated communication as a force in language change.

Visible language 18, 2 (1984), 118.

[17] BARTH, L., FABRIKANT, S., KOBOUROV, S. G., LUBIW, A., NÖLLENBURG, M., OKAMOTO, Y., PUPYREV, S., SQUARCELLA, C., UECKERDT, T., AND WOLFF, A.

Semantic word cloud representations: Hardness and approximation algorithms.

In *Latin American Symposium on Theoretical Informatics* (2014), Springer, pp. 514–525.

[18] BAYES, T.

An essay towards solving a problem in the doctrine of chances. 1763.

MD computing: computers in medical practice 8, 3 (1991), 157.

[19] BEAUFORT, R., ROEKHAUT, S., COUGNON, L.-A., AND FAIRON, C.

A hybrid rule/model-based finite-state framework for normalizing sms messages.

In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (2010), Association for Computational Linguistics, pp. 770–779.

[20] BEKOS, M. A., VAN DIJK, T. C., FINK, M., KINDERMANN, P., KOBOUROV, S., PUPYREV, S., SPOERHASE, J., AND WOLFF, A.

Improved approximation algorithms for box contact representations.

Algorithmica 77, 3 (2017), 902–920.

[21] BERGSMA, S., AND VAN DURME, B.

Using conceptual class attributes to characterize social media users.

In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (2013), vol. 1, pp. 710–720.

[22] BIBER, D., CONRAD, S., AND REPPEN, R.

Corpus linguistics: Investigating language structure and use.

Cambridge University Press, 1998.

[23] BISSIAS, G., LEVINE, B., LIBERATORE, M., LYNN, B., MOORE, J., WALLACH, H., AND WOLAK, J.

Characterization of contact offenders and child exploitation material trafficking on five peer-to-peer networks.

Child abuse & neglect 52 (2016), 185–199.

[24] BLEI, D. M., NG, A. Y., AND JORDAN, M. I.

Latent dirichlet allocation.

Journal of machine Learning research 3, Jan (2003), 993–1022.

[25] BOGDANOVA, D., ROSSO, P., AND SOLORIO, T.

Exploring high-level features for detecting cyberpedophilia.

Computer speech & language 28, 1 (2014), 108–120.

[26] BOLLEGALA, D., MATSUO, Y., AND ISHIZUKA, M.

Measuring semantic similarity between words using web search engines.

www 7 (2007), 757–766.

[27] BOSCH, A. V. D., BUSSER, B., CANISIUS, S., AND DAELEMANS, W.

An efficient memory-based morphosyntactic tagger and parser for Dutch.

LOT Occasional Series 7 (2007), 191–206.

[28] BOURKE, M. L., FRAGOMELI, L., DETAR, P. J., SULLIVAN, M. A., MEYLE, E., AND O'RIORDAN, M.

The use of tactical polygraph with sex offenders.

Journal of Sexual Aggression 21, 3 (2015), 354–367.

[29] BOURKE, M. L., AND HERNANDEZ, A. E.

The 'butner study'redux: A report of the incidence of hands-on child victimization by child pornography offenders.

Journal of Family Violence 24, 3 (2009), 183.

[30] BREIMAN, F.

Olshen, and stone.

Classification and Regression trees (1984).

[31] BRENNAN, M., AFROZ, S., AND GREENSTADT, R.

Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity.

ACM Transactions on Information and System Security (TISSEC) 15, 3 (2012), 12.

[32] BRENNAN, M. R., AND GREENSTADT, R.

Practical attacks against authorship recognition techniques.

In *IAAI* (2009).

[33] BRESLOW, N. E., AND CLAYTON, D. G.

Approximate inference in generalized linear mixed models.

Journal of the American statistical Association 88, 421 (1993), 9–25.

[34] BURGER, J. D., HENDERSON, J., KIM, G., AND ZARRELLA, G.

Discriminating gender on Twitter.

In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (2011), Association for Computational Linguistics, pp. 1301–1309.

[35] BURROWS, J.

All the way through: testing for authorship in different frequency strata.

Literary and Linguistic Computing 22, 1 (2006), 27–47.

[36] CALLE-MARTÍN, J., AND MIRANDA-GARCÍA, A.

Stylometry and authorship attribution: Introduction to the special issue.

English Studies 93, 3 (2012), 251–258.

[37] CAVNAR, W. B., TRENKLE, J. M., ET AL.

N-gram-based text categorization.

Ann Arbor MI 48113, 2 (1994), 161–175.

[38] CHAMBERS, J. K.

Sociolinguistic theory.

Blackwell, 1995.

[39] CHAMBERS, J. K., AND SCHILLING, N.

The handbook of language variation and change, vol. 129.

John Wiley & Sons, 2013.

[40] CHASKI, C. E.

Who's at the keyboard? Authorship attribution in digital evidence investigations.

International journal of digital evidence 4, 1 (2005), 1–13.

[41] CHAWLA, N. V., JAPKOWICZ, N., AND KOTCZ, A.

Special issue on learning from imbalanced data sets.

ACM Sigkdd Explorations Newsletter 6, 1 (2004), 1–6.

[42] CHESHIRE, J.

Sex and gender in variationist research.

The handbook of language variation and change (2002), 423–443.

[43] CHUNG, C., AND PENNEBAKER, J. W.

The psychological functions of function words.

Social communication (2007), 343–359.

[44] COOPER, A.

Sexuality and the Internet: Surfing into the new millennium.

CyberPsychology & Behavior 1, 2 (1998), 187–193.

[45] CORTES, C., AND VAPNIK, V.

Support vector machine.

Machine learning 20, 3 (1995), 273–297.

[46] COVER, T. M., AND THOMAS, J. A.

Elements of information theory.

John Wiley & Sons, 2012.

[47] CPS.

Child protection system. P2P monitoring software developed at TLO. http://www.tlo.com/.

[48] CRAVEN, S., BROWN, S., AND GILCHRIST, E.

Sexual grooming of children: Review of literature and theoretical considerations.

Journal of Sexual Aggression 12, 3 (2006), 287–299.

[49] CRYSTAL, D.

Language and the Internet.

Cambridge, CUP (2001).

[50] DAELEMANS, W.

Explanation in computational stylometry.

In *International Conference on Intelligent Text Processing and Computational Linguistics* (2013), Springer, pp. 451–462.

[51] DAI, L.

An Ising-based approach for tracking illegal P2P content distributors.

PhD thesis, Iowa State University, 2010.

[52] DINGLEDINE, R., MATHEWSON, N., AND SYVERSON, P.

Tor: The second-generation onion router.

Tech. rep., Naval Research Lab Washington DC, 2004.

[53] DINU, L. P., AND NISIOI, S.

Authorial studies using ranked lexical features.

In *COLING (Demos)* (2012), pp. 125–130.

[54] DOBSON, M.

The Making of the National Poet: Shakespeare, Adaptation and Authorship, 1660-1769: Shakespeare, Adaptation and Authorship, 1660-1769.

Clarendon Press, 1992.

[55] DOWNES, W.

Language and society.

Cambridge University Press, 1998.

[56] DUNNING, T.

Statistical identification of language.

Computing Research Laboratory, New Mexico State University, 1994.

[57] ECPAT INTERNATIONAL.

Towards a global indicator on unidentified victims in child sexual exploitation material.

`http://www.ecpat.org/wp-content/uploads/2018/03/TOWARDS-A-GLOBAL-INDICATOR-ON-UNIDENTIFIED-VICTIMS-IN-CHILD-SEXUAL-EXPLOITATION-MATERIAL-Summary-Report.pdf`, March 2018.

[58] EDWARDS, M., PEERSMAN, C., AND RASHID, A.

Scamming the scammers: towards automatic detection of persuasion in advance fee frauds.

In *Proceedings of the 26th International Conference on World Wide Web Companion* (2017), International World Wide Web Conferences Steering Committee, pp. 1291–1299.

[59] EDWARDS, M., AND RASHID, A.

Collaborative filtering as an investigative tool for peer-to-peer filesharing networks.

SCIENCE 1, 2 (2012), pp–67.

[60] EISENSTEIN, J., O'CONNOR, B., SMITH, N. A., AND XING, E. P.

A latent variable model for geographic lexical variation.

In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (2010), Association for Computational Linguistics, pp. 1277–1287.

[61] ELZINGA, P., WOLFF, K. E., AND POELMANS, J.

Analyzing chat conversations of pedophiles with temporal relational semantic systems.

In *Intelligence and Security Informatics Conference (EISIC), 2012 European* (2012), IEEE, pp. 242–249.

[62] ERIKSSON, G., AND KARLGREN, J.

Features for modelling characteristics of conversations: Notebook for PAN at CLEF 2012.
In *CLEF 2012 Evaluation Labs and Workshop, Rome, Italy, 17-20 September 2012* (2012).

[63] FAN, R.-E., CHANG, K.-W., HSIEH, C.-J., WANG, X.-R., AND LIN, C.-J.
Liblinear: A library for large linear classification.
Journal of machine learning research 9, Aug (2008), 1871–1874.

[64] FELDMAN, R., AND SANGER, J.
The text mining handbook: advanced approaches in analyzing unstructured data.
Cambridge university press, 2007.

[65] FILIPPOVA, K.
User demographics and language in an implicit social network.
In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (2012), Association for Computational Linguistics, pp. 1478–1488.

[66] FINK, C., KOPECKY, J., AND MORAWSKI, M.
Inferring gender from the content of tweets: A region specific example.
In *ICWSM* (2012).

[67] FORMAN, G.
An extensive empirical study of feature selection metrics for text classification.
Journal of machine learning research 3, Mar (2003), 1289–1305.

[68] FRANTZESKOU, G., STAMATATOS, E., GRITZALIS, S., CHASKI, C. E., AND HOWALD, B. S.
Identifying authorship by byte-level n-grams: The source code author profile (scap) method.
International Journal of Digital Evidence 6, 1 (2007), 1–18.

[69] FRIEDMAN, J., HASTIE, T., AND TIBSHIRANI, R.
The elements of statistical learning, vol. 1.
Springer series in statistics New York, 2001.

[70] FROMKIN, V., RODMAN, R., AND HYAMS, N.

An introduction to language.

Cengage Learning, 2013.

[71] GÁLVEZ, R. H., AND GRAVANO, A.

Assessing the usefulness of online message board mining in automatic stock prediction systems.

Journal of Computational Science 19 (2017), 43–56.

[72] GEERAERTS, D.

Een zondagspak? Het Nederlands in Vlaanderen: gedrag, beleid, attitudes.

Ons erfdeel 44, 3 (2001), 337–343.

[73] GOSWAMI, S., SARKAR, S., AND RUSTAGI, M.

Stylometric analysis of bloggers' age and gender.

In *Third International AAAI Conference on Weblogs and Social Media* (2009).

[74] GOTTSCHALK, P.

A dark side of computing and information sciences: characteristics of online groomers.

Journal of Emerging Trends in Computing and Information Sciences 2, 9 (2011), 447–455.

[75] GRONDELAERS, S., AND VAN HOUT, R.

The standard language situation in the low countries: Top-down and bottom-up variations on a diaglossic theme.

Journal of Germanic Linguistics 23, 3 (2011), 199–243.

[76] GUMMADI, K. P., DUNN, R. J., SAROIU, S., GRIBBLE, S. D., LEVY, H. M., AND ZAHORJAN, J.

Measurement, modeling, and analysis of a peer-to-peer file-sharing workload.

ACM SIGOPS Operating Systems Review 37, 5 (2003), 314–329.

[77] GUTHRIE, D., ALLISON, B., LIU, W., GUTHRIE, L., AND WILKS, Y.

A closer look at skip-gram modelling.

In *Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC-2006)* (2006), sn, pp. 1–4.

[78] HALL, R. C., AND HALL, R. C.

A profile of pedophilia: definition, characteristics of offenders, recidivism, treatment outcomes, and forensic issues.

In *Mayo Clinic Proceedings* (2007), vol. 82, Elsevier, pp. 457–471.

[79] HARRELL, F. E.

Regression modeling strategies, with applications to linear models, survival analysis and logistic regression.

Springer (2001).

[80] HEAP, B., BAIN, M., WOBCKE, W., KRZYWICKI, A., AND SCHMEIDL, S.

Word vector enrichment of low frequency words in the bag-of-words model for short text multi-class classification problems.

arXiv preprint arXiv:1709.05778 (2017).

[81] HECK, L. P., KONIG, Y., SÖNMEZ, M. K., AND WEINTRAUB, M.

Robustness to telephone handset distortion in speaker recognition by discriminative feature design.

Speech Communication 31, 2-3 (2000), 181–192.

[82] HERRING, S., STEIN, D., AND VIRTANEN, T.

Pragmatics of computer-mediated communication, vol. 9.

Walter de Gruyter, 2013.

[83] HIDALGO, J. M. G., AND DÍAZ, A. A. C.

Combining predation heuristics and chat-like features in sexual predator identification.

In *CLEF (Online Working Notes/Labs/Workshop)* (2012).

[84] HIRST, G., AND FEIGUINA, O.

Bigrams of syntactic labels for authorship discrimination of short texts.

Literary and Linguistic Computing 22, 4 (2007), 405–417.

[85] HOFFMAN, M., BACH, F. R., AND BLEI, D. M.

Online learning for latent dirichlet allocation.

In *advances in neural information processing systems* (2010), pp. 856–864.

[86] HOLMES, J.

An introduction to sociolinguistics.

Routledge, 2013.

[87] HSIEH, C.-J., CHANG, K.-W., LIN, C.-J., KEERTHI, S. S., AND SUNDARARAJAN, S.

A dual coordinate descent method for large-scale linear SVM.

In *Proceedings of the 25th international conference on Machine learning* (2008), ACM, pp. 408–415.

[88] HU, X., SUN, N., ZHANG, C., AND CHUA, T.-S.

Exploiting internal and external semantics for the clustering of short texts using world knowledge.

In *Proceedings of the 18th ACM conference on Information and knowledge management* (2009), ACM, pp. 919–928.

[89] HUGHES, D., RAYSON, P., WALKERDINE, J., LEE, K., GREENWOOD, P., RASHID, A., MAY-CHAHAL, C., AND BRENNAN, M.

Supporting law enforcement in digital communities through natural language analysis.

Computational Forensics (2008), 122–134.

[90] HUGHES, D., WALKERDINE, J., COULSON, G., AND GIBSON, S.

Peer-to-peer: Is deviant behavior the norm on P2P file-sharing networks?

IEEE distributed systems online 7, 2 (2006).

[91] HURLEY, R., PRUSTY, S., SOROUSH, H., WALLS, R. J., ALBRECHT, J., CECCHET, E., LEVINE, B. N., LIBERATORE, M., LYNN, B., AND WOLAK, J.

Measurement and analysis of child pornography trafficking on P2P networks.

In *Proceedings of the 22nd international conference on World Wide Web* (2013), ACM, pp. 631–642.

[92] INCHES, G., AND CRESTANI, F.

Overview of the international sexual predator identification competition at PAN-2012.

In *CLEF (Online working notes/labs/workshop)* (2012), vol. 30.

[93] INTERNATIONAL ASSOCIATION OF INTERNET HOTLINES.

INHOPE Annual Report 2017.

http://www.inhope.org/Libraries/Annual_reports/INHOPE_Annual_Report_2017.sflb.ashx?, December 2017.

[94] IRONS, A., AND LALLIE, H. S.

Digital forensics to intelligent forensics.

Future Internet 6, 3 (2014), 584–596.

[95] JAIN, A. K.

Data clustering: 50 years beyond k-means.

Pattern recognition letters 31, 8 (2010), 651–666.

[96] JENEY, P.

Combatting child sexual abuse online.

http://www.europarl.europa.eu/RegData/etudes/STUD/2015/536481/IPOL_STU(2015)536481_EN.pdf, 2015.

[97] JOFFRES, K., BOUCHARD, M., FRANK, R., AND WESTLAKE, B.

Strategies to disrupt online child pornography networks.

In *Intelligence and Security Informatics Conference (EISIC), 2011 European* (2011), IEEE, pp. 163–170.

[98] JUOLA, P.

Authorship attribution.

Foundations and Trends® in Information Retrieval 1, 3 (2008), 233–334.

[99] JUOLA, P.

Detecting stylistic deception.

In *Proceedings of the Workshop on Computational Approaches to Deception Detection* (2012), Association for Computational Linguistics, pp. 91–96.

[100] JUOLA, P., AND VESCOVI, D.

Empirical evaluation of authorship obfuscation using JGAAP.

In *Proceedings of the 3rd ACM workshop on Artificial Intelligence and Security* (2010), ACM, pp. 14–18.

[101] JUOLA, P., AND VESCOVI, D.

Analyzing stylometric approaches to author obfuscation.

Advances in Digital Forensics VII (2011), 115–125.

[102] KACMARCIK, G., AND GAMON, M.

Obfuscating document stylometry to preserve author anonymity.

In *Proceedings of the COLING/ACL on Main conference poster sessions* (2006), Association for Computational Linguistics, pp. 444–451.

[103] KELLER, F., LAPATA, M., AND OURIOUPINA, O.

Using the web to overcome data sparseness.

In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10* (2002), Association for Computational Linguistics, pp. 230–237.

[104] KEMP, S.

Global digital report 2018.

`https://wearesocial.com/uk/blog/2018/01/global-digital-report-2018`, 2018.

[105] KERN, R., KLAMPFL, S., AND ZECHNER, M.

Vote/veto classification, ensemble clustering and sequence classification for author identification.

In *CLEF (Online Working Notes/Labs/Workshop)* (2012).

[106] KESTEMONT, M., LUYCKX, K., DAELEMANS, W., AND CROMBEZ, T.

Cross-genre authorship verification using unmasking.

English Studies 93, 3 (2012), 340–356.

[107] KESTEMONT, M., PEERSMAN, C., DE DECKER, B., DE PAUW, G., LUYCKX, K., MORANTE, R., VAASSEN, F., VAN DE LOO, J., AND DAELEMANS, W.
The Netlog Corpus. a resource for the study of Flemish Dutch Internet language.
In *LREC* (2012), Citeseer, pp. 1569–1572.

[108] KEUNE, K.
Explaining register and sociolinguistic variation in the lexicon: Corpus studies on Dutch.
PhD thesis, Radboud Universiteit Nijmegen, 2012.

[109] KIESLER, S., SIEGEL, J., AND MCGUIRE, T. W.
Social psychological aspects of computer-mediated communication.
American psychologist 39, 10 (1984), 1123.

[110] KINGMA, D. P., AND BA, J.
Adam: A method for stochastic optimization.
arXiv preprint arXiv:1412.6980 (2014).

[111] KONTOSTATHIS, A., GARRON, A., REYNOLDS, K., WEST, W., AND EDWARDS, L.
Identifying predators using ChatCoder 2.0.
In *CLEF (Online Working Notes/Labs/Workshop)* (2012).

[112] KOPPEL, M., ARGAMON, S., AND SHIMONI, A. R.
Automatically categorizing written texts by author gender.
Literary and Linguistic Computing 17, 4 (2002), 401–412.

[113] KOPPEL, M., AND SCHLER, J.
Exploiting stylistic idiosyncrasies for authorship attribution.
In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis* (2003), vol. 69, p. 72.

[114] KOPPEL, M., SCHLER, J., AND ARGAMON, S.
Computational methods in authorship attribution.
Journal of the Association for Information Science and Technology 60, 1 (2009), 9–26.

[115] KOPPEL, M., SCHLER, J., AND ARGAMON, S.

Authorship attribution in the wild.

Language Resources and Evaluation 45, 1 (2011), 83–94.

[116] KOPPEL, M., SCHLER, J., AND BONCHEK-DOKOW, E.

Measuring differentiability: Unmasking pseudonymous authors.

Journal of Machine Learning Research 8, Jun (2007), 1261–1276.

[117] KOTSIANTIS, S. B., ZAHARAKIS, I., AND PINTELAS, P.

Supervised machine learning: A review of classification techniques.

Emerging artificial intelligence applications in computer engineering 160 (2007), 3–24.

[118] KRASKOV, A., STÖGBAUER, H., AND GRASSBERGER, P.

Estimating mutual information.

Physical review E 69, 6 (2004), 066138.

[119] LABOV, W.

Sociolinguistic patterns.

No. 4. University of Pennsylvania Press, 1972.

[120] LABOV, W.

The intersection of sex and social class in the course of linguistic change.

Language variation and change 2, 2 (1990), 205–254.

[121] LABOV, W.

Principles of linguistic change Volume 1: Internal Factors.

Blackwell, 1994.

[122] LABOV, W.

Principles of linguistic change Volume 2: Social Factors.

Blackwell, 2001.

[123] LANNING, K.

Child molesters: A behavioral analysis for professionals investigating the sexual exploitation of children.

National Center for Missing & Exploited Children with Office of Juvenile Justice and Delinquency Prevention, 2010.

[124] LATAPY, M., MAGNIEN, C., AND FOURNIER, R.

Quantifying paedophile activity in a large P2P system.

Information Processing & Management 49, 1 (2013), 248–263.

[125] LEWIS, D. D.

Naive (bayes) at forty: The independence assumption in information retrieval.

In *European conference on machine learning* (1998), Springer, pp. 4–15.

[126] LI, Y., WANG, J.-L., TIAN, Z.-H., LU, T.-B., AND YOUNG, C.

Building lightweight intrusion detection system using wrapper-based feature selection mechanisms.

Computers & Security 28, 6 (2009), 466–475.

[127] LIBERATORE, M., ERDELY, R., KERLE, T., LEVINE, B. N., AND SHIELDS, C.

Forensic investigation of peer-to-peer file sharing networks.

digital investigation 7 (2010), S95–S103.

[128] LIBERATORE, M., LEVINE, B. N., AND SHIELDS, C.

Strengthening forensic investigations of child pornography on P2P networks.

In *Proceedings of the 6th International COnference* (2010), ACM, p. 19.

[129] LIU, M., HAFFARI, G., BUNTINE, W., AND ANANDA-RAJAH, M.

Leveraging linguistic resources for improving neural text classification.

In *Proceedings of the Australasian Language Technology Association Workshop 2017* (2017), pp. 34–42.

[130] LIVINGSTONE, S., HADDON, L., GÖRZIG, A., AND ÓLAFSSON, K.

Risks and safety on the Internet: the perspective of European children: full findings and policy implications from the EU Kids Online survey of 9–16 year olds and their parents in 25 countries.

`http://eprints.lse.ac.uk/33731/1/Risks%20and%20safety%20on%20the%20internet%28lsero%29.pdf`, 2011.

[131] LUYCKX, K.

Scalability issues in authorship attribution.

PhD thesis, Universiteit Antwerpen, 2011.

[132] LUYCKX, K., AND DAELEMANS, W.

The effect of author set size and data size in authorship attribution.

Literary and linguistic Computing 26, 1 (2011), 35–55.

[133] MARON, M. E.

Automatic indexing: an experimental inquiry.

Journal of the ACM (JACM) 8, 3 (1961), 404–417.

[134] MARON, M. E., AND KUHNS, J. L.

On relevance, probabilistic indexing and information retrieval.

Journal of the ACM (JACM) 7, 3 (1960), 216–244.

[135] MCGHEE, I., BAYZICK, J., KONTOSTATHIS, A., EDWARDS, L., MCBRIDE, A., AND JAKUBOWSKI, E.

Learning to identify Internet sexual predation.

International Journal of Electronic Commerce 15, 3 (2011), 103–122.

[136] MCMILLAN, J. R., CLIFTON, A. K., MCGRATH, D., AND GALE, W. S.

Women's language: Uncertainty or interpersonal sensitivity and emotionality?

Sex Roles 3, 6 (1977), 545–559.

[137] MEHL, M. R., AND PENNEBAKER, J. W.

The sounds of social life: a psychometric analysis of students' daily social environments and natural conversations.

Journal of personality and social psychology 84, 4 (2003), 857.

[138] MENASCHE, D. S., ROCHA, D. A., ANTONIO, A., LI, B., TOWSLEY, D., AND VENKATARAMANI, A.

Content availability and bundling in swarming systems.

IEEE/ACM Transactions on Networking (TON) 21, 2 (2013), 580–593.

[139] MENG, W., LANFEN, L., JING, W., PENGHUA, Y., JIAOLONG, L., AND FEI, X.

Improving short text classification using public search engines.

In *International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making* (2013), Springer, pp. 157–166.

[140] MICHALOPOULOS, D., AND MAVRIDIS, I.

Utilizing document classification for grooming attack recognition.

In *Computers and Communications (ISCC), 2011 IEEE Symposium on* (2011), IEEE, pp. 864–869.

[141] MICROSOFT.

New technology fights child porn by tracking its "photodna". http://www.microsoft.com/en-us/news/features/2009/dec09/12-15photodna.mspx, 2009.

[142] MIDDLETON, D.

Internet sex offenders.

Assessment and treatment of sex offenders: A handbook (2009), 199–215.

[143] MILROY, J., MILROY, L., HARTLEY, S., AND WALSHAW, D.

Glottal stops and tyneside glottalization: Competing patterns of variation and change in British English.

Language Variation and Change 6, 3 (1994), 327–357.

[144] MORRIS, C., AND HIRST, G.

Identifying sexual predators by SVM classification with lexical and behavioral features.

In *CLEF (Online Working Notes/Labs/Workshop)* (2012), vol. 12, p. 29.

[145] MOSTELLER, F., AND WALLACE, D.
Inference and disputed authorship: The Federalist.
Addison-Wesley, 1964.

[146] MUKHERJEE, A., AND LIU, B.
Improving gender classification of blog authors.
In *Proceedings of the 2010 conference on Empirical Methods in natural Language Processing* (2010), Association for Computational Linguistics, pp. 207–217.

[147] MULAC, A., SEIBOLD, D. R., AND FARRIS, J. L.
Female and male managers' and professionals' criticism giving: Differences in language use and effects.
Journal of Language and Social Psychology 19, 4 (2000), 389–415.

[148] NERBONNE, J.
The secret life of pronouns. What our words say about us.
Literary and Linguistic Computing 29, 1 (2014), 139–142.

[149] NEWCOMBE, R. G.
Two-sided confidence intervals for the single proportion: comparison of seven methods.
Statistics in medicine 17, 8 (1998), 857–872.

[150] NEWMAN, M. L., GROOM, C. J., HANDELMAN, L. D., AND PENNEBAKER, J. W.
Gender differences in language use: An analysis of 14,000 text samples.
Discourse Processes 45, 3 (2008), 211–236.

[151] NGUYEN, D., DOĞRUÖZ, A. S., ROSÉ, C. P., AND DE JONG, F.
Computational sociolinguistics: A survey.
Computational Linguistics 42, 3 (2016), 537–593.

[152] NGUYEN, D., GRAVEL, R., TRIESCHNIGG, D., AND MEDER, T.
'How old do you think I am?' A study of language and age in Twitter.
In *ICWSM* (2013).

[153] NGUYEN, D., SMITH, N. A., AND ROSÉ, C. P.

Author age prediction from text using linear regression.

In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (2011), Association for Computational Linguistics, pp. 115–123.

[154] NOWSON, S.

Identifying more bloggers: Towards large scale personality classifiation of personal weblogs.

`http://nowson.com/papers/NowOberICWSM07.pdf`.

[155] OAKES, M.

Ant colony optimisation for stylometry: The Federalist papers.

In *Proceedings of the 5th International Conference on Recent Advances in Soft Computing* (2004), pp. 86–91.

[156] PANCHENKO, A., BEAUFORT, R., AND FAIRON, C.

Detection of child sexual abuse media on P2P networks: Normalization and classification of associated filenames.

In *Proceedings of the LREC Workshop on Language Resources for Public Security Applications* (2012).

[157] PARAPAR, J., LOSADA, D. E., AND BARREIRO, A.

A learning-based approach for the identification of sexual predators in chat logs.

In *CLEF (Online Working Notes/Labs/Workshop)* (2012).

[158] PASQUALE, L., ALRAJEH, D., PEERSMAN, C., TUN, T., NUSEIBEH, B., AND RASHID, A.

Towards forensic-ready software systems.

In *Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results* (2018), ACM, pp. 9–12.

[159] PEERSMAN, C., DAELEMANS, W., AND VAN VAERENBERGH, L.

Predicting age and gender in online social networks.

In *Proceedings of the 3rd international workshop on Search and mining user-generated contents* (2011), ACM, pp. 37–44.

[160] PEERSMAN, C., DAELEMANS, W., VANDEKERCKHOVE, R., VANDEKERCKHOVE, B., AND VAN VAERENBERGH, L.

The effects of age, gender and region on non-standard linguistic variation in online social networks.

arXiv preprint arXiv:1601.02431 (2016).

[161] PEERSMAN, C., SCHULZE, C., RASHID, A., BRENNAN, M., AND FISCHER, C.

iCOP: Automatically identifying new child abuse media in P2P networks.

In *Security and Privacy Workshops (SPW), 2014 IEEE* (2014), IEEE, pp. 124–131.

[162] PEERSMAN, C., SCHULZE, C., RASHID, A., BRENNAN, M., AND FISCHER, C.

iCOP: Live forensics to reveal previously unknown criminal media on P2P networks.

Digital Investigation 18 (2016), 50–64.

[163] PEERSMAN, C., VAASSEN, F., VAN ASCH, V., AND DAELEMANS, W.

Conversation level constraints on pedophile detection in chat rooms.

In *CLEF (Online Working Notes/Labs/Workshop)* (2012).

[164] PENDAR, N.

Toward spotting the pedophile telling victim from predator in text chats.

In *Semantic Computing, 2007. ICSC 2007. International Conference on* (2007), IEEE, pp. 235–241.

[165] PENNEBAKER, J. W., BOYD, R. L., JORDAN, K., AND BLACKBURN, K.

The development and psychometric properties of LIWC 2015.

Tech. rep., The University of Texas at Austin, 2015.

[166] PENNEBAKER, J. W., FRANCIS, M. E., AND BOOTH, R. J.

Linguistic inquiry and word count: LIWC 2001.

Mahway: Lawrence Erlbaum Associates 71, 2001 (2001), 2001.

[167] PENNEBAKER, J. W., AND KING, L. A.

Linguistic styles: language use as an individual difference.

Journal of personality and social psychology 77, 6 (1999), 1296.

[168] PLEVOETS, K.

Tussen spreek-en standaardtaal. Een corpusgebaseerd onderzoek naar de situationele, regionale en sociale verspreiding van enkele morfosyntactische verschijnselen uit het gesproken Belgisch-Nederlands.

PhD thesis, Katholieke Universiteit Leuven, 2008.

[169] POPESCU, M., AND GROZEA, C.

Kernel methods and string kernels for authorship analysis.

In *CLEF (Online Working Notes/Labs/Workshop)* (2012).

[170] PRICHARD, J., WATTERS, P. A., AND SPIRANOVIC, C.

Internet subcultures and pathways to the use of child pornography.

Computer Law & Security Review 27, 6 (2011), 585–600.

[171] QUAYLE, E., HOLLAND, G., LINEHAN, C., AND TAYLOR, M.

The Internet and offending behaviour: A case study.

Journal of Sexual Aggression 6, 1-2 (2000), 78–96.

[172] RAHMANMIAH, M. W., YEARWOOD, J., AND KULKARNI, S.

Detection of child exploiting chats from a mixed chat dataset as text classification task.

In *Proceedings of the Australian Language Technology Association Workshop* (2011), pp. 157–165.

[173] RANGEL, F., ROSSO, P., CHUGUR, I., POTTHAST, M., TRENKMANN, M., STEIN, B., VERHOEVEN, B., AND DAELEMANS, W.

Overview of the 2nd author profiling task at PAN 2014.

In *CLEF 2014 Evaluation Labs and Workshop Working Notes Papers, Sheffield, UK, 2014* (2014), pp. 1–30.

[174] RANGEL, F., ROSSO, P., KOPPEL, M., STAMATATOS, E., AND INCHES, G.
Overview of the author profiling task at PAN 2013.
In *CLEF Conference on Multilingual and Multimodal Information Access Evaluation* (2013), CELCT, pp. 352–365.

[175] RANGEL, F., ROSSO, P., POTTHAST, M., STEIN, B., AND DAELEMANS, W.
Overview of the 3rd author profiling task at PAN 2015.
In *CLEF* (2015), sn, p. 2015.

[176] RAO, D., YAROWSKY, D., SHREEVATS, A., AND GUPTA, M.
Classifying latent user attributes in Twitter.
In *Proceedings of the 2nd international workshop on Search and mining user-generated contents* (2010), ACM, pp. 37–44.

[177] RASHID, A., BARON, A., RAYSON, P., MAY-CHAHAL, C., GREENWOOD, P., AND WALKERDINE, J.
Who am I? Analyzing digital personas in cybercrime investigations.
Computer 46, 4 (2013), 54–61.

[178] REED, M. G., SYVERSON, P. F., AND GOLDSCHLAG, D. M.
Anonymous connections and onion routing.
IEEE Journal on Selected areas in Communications 16, 4 (1998), 482–494.

[179] REHUREK, R.
Gensim: Topic modelling for humans (2017).
`https://radimrehurek.com/gensim/models/keyedvectors.html`.

[180] ROCHA, A., SCHEIRER, W. J., FORSTALL, C. W., CAVALCANTE, T., THEOPHILO, A., SHEN, B., CARVALHO, A. R., AND STAMATATOS, E.
Authorship attribution for social media forensics.
IEEE Transactions on Information Forensics and Security 12, 1 (2017), 5–33.

[181] ROSS, B. C.

Mutual information between discrete and continuous data sets.

PloS one 9, 2 (2014), e87357.

[182] ROUSE, J.

Child sexual offenders methodologies in southeast asia – from live streaming to social media hunting.

`https://sites.google.com/view/ocseconference2017/speakers`, 2017.

Presented at the UN conference "Effective Responses to Online Child Sexual Exploitation in Southeast Asia" in Bangkok, Thailand.

[183] RUMELHART, D. E., HINTON, G. E., AND WILLIAMS, R. J.

Learning internal representations by error propagation.

Tech. rep., California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[184] SANDERSON, C., AND GUENTER, S.

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* (2006), Association for Computational Linguistics, pp. 482–491.

[185] SARAWGI, R., GAJULAPALLI, K., AND CHOI, Y.

Gender attribution: tracing stylometric evidence beyond topic and genre.

In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning* (2011), Association for Computational Linguistics, pp. 78–86.

[186] SCHLER, J., KOPPEL, M., ARGAMON, S., AND PENNEBAKER, J. W.

Effects of age and gender on blogging.

In *AAAI spring symposium: Computational approaches to analyzing weblogs* (2006), vol. 6, pp. 199–205.

[187] SCIKIT-LEARN.

Scikit-learn user guide.

`https://scikit-learn.org/stable/_downloads/scikit-learn-docs.pdf`.

[188] SEBASTIANI, F.

Machine learning in automated text categorization.

ACM computing surveys (CSUR) 34, 1 (2002), 1–47.

[189] SEIFFERT, C., KHOSHGOFTAAR, T. M., VAN HULSE, J., AND FOLLECO, A.

An empirical study of the classification performance of learners on imbalanced and noisy software quality data.

Information Sciences 259 (2014), 571–595.

[190] SETO, M. C., KARL HANSON, R., AND BABCHISHIN, K. M.

Contact sexual offending by men with online sexual offenses.

Sexual Abuse 23, 1 (2011), 124–145.

[191] SIEBENHAAR, B.

Code choice and code-switching in Swiss-German Internet relay chat rooms.

Journal of Sociolinguistics 10, 4 (2006), 481–506.

[192] SNYDER, H. N.

Sexual assault of young children as reported to law enforcement: victim, incident, and offender characteristics.

DIANE Publishing, 2010.

[193] SONG, D., BRUZA, P., HUANG, Z., AND LAU, R. Y.

Classifying document titles based on information inference.

In *International Symposium on Methodologies for Intelligent Systems* (2003), Springer, pp. 297–306.

[194] SONG, G., YE, Y., DU, X., HUANG, X., AND BIE, S.

Short text classification: A survey.

Journal of Multimedia 9, 5 (2014).

[195] SRIRAM, B., FUHRY, D., DEMIR, E., FERHATOSMANOGLU, H., AND DEMIRBAS, M.

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* (2010), ACM, pp. 841–842.

[196] STUTZBACH, D., AND REJAIE, R.

Understanding churn in peer-to-peer networks.

In *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement* (2006), ACM, pp. 189–202.

[197] TAELDEMAN, J.

Zich stabiliserende grammaticale kenmerken in Vlaamse tussentaal.

Taal en Tongval 60, 1 (2008), 26–50.

[198] TAGLIAMONTE, S.

Variationist sociolinguistics: Change, observation, interpretation, vol. 40.

John Wiley & Sons, 2012.

[199] THURLOW, C., AND POFF, M.

Text messaging.

Pragmatics of computer-mediated communication 94 (2013).

[200] TONG, S., AND KOLLER, D.

Support vector machine active learning with applications to text classification.

Journal of machine learning research 2, Nov (2001), 45–66.

[201] TROTTA, J.

Time, tense and aspect in nonstandard english: an overview.

In *Tid och tidsförhållanden i olika språk*. 2011, pp. 139–158.

[202] TRUDGILL, P.

On dialect: Social and geographical perspectives.

Wiley-Blackwell, 1983.

[203] TSOUMAKAS, G., AND KATAKIS, I.

Multi-label classification: An overview.

International Journal of Data Warehousing and Mining 3, 3 (2006).

[204] VAN DE LOO, J., DE PAUW, G., AND DAELEMANS, W.
Text-based age and gender prediction for online safety monitoring.
Comput. Linguistics Netherlands 5, 1 (2016), 46–60.

[205] VAN HALTEREN, H.
Linguistic profiling for author recognition and verification.
In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics* (2004), Association for Computational Linguistics, p. 199.

[206] VAN HALTEREN, H., BAAYEN, H., TWEEDIE, F., HAVERKORT, M., AND NEIJT, A.
New machine learning methods demonstrate the existence of a human stylome.
Journal of Quantitative Linguistics 12, 1 (2005), 65–77.

[207] VAN HOOF, S., ET AL.
Feiten en fictie-taalvariatie in Vlaamse televisiereeksen vroeger en nu.
Nederlandse taalkunde 18, 1 (2013), 35–64.

[208] VAN KEYMEULEN, J.
Een verkennend taalgeografisch onderzoek naar dialectverlies in Nederlandstalig België.
Taal en Tongval 6 (1993), 120–135.

[209] VAN WISSEN, L., AND BOOT, P.
An electronic translation of the LIWC dictionary into Dutch.
In *Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference* (2017), Lexical Computing, pp. 703–715.

[210] VANDEKERCKHOVE, R.
Dialect loss and dialect vitality in Flanders.
International Journal of the Sociology of Language 196 | 197 (2009), 73–97.

[211] VANDEKERCKHOVE, R., AND NOBELS, J.

Code eclecticism: Linguistic variation and code alternation in the chat language of Flemish teenagers.

Journal of Sociolinguistics 14, 5 (2010), 657–677.

[212] VERGARA, J. R., AND ESTÉVEZ, P. A.

A review of feature selection methods based on mutual information.

Neural computing and applications 24, 1 (2014), 175–186.

[213] VILLATORO-TELLO, E., JUÁREZ-GONZÁLEZ, A., ESCALANTE, H. J., MONTES-Y GÓMEZ, M., AND PINEDA, L. V.

A two-step approach for effective detection of misbehaving users in chats.

In *CLEF (Online Working Notes/Labs/Workshop)* (2012).

[214] WANG, F., WANG, Z., LI, Z., AND WEN, J.-R.

Concept-based short text classification and ranking.

In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (2014), ACM, pp. 1069–1078.

[215] WARD JR, J. H.

Hierarchical grouping to optimize an objective function.

Journal of the American statistical association 58, 301 (1963), 236–244.

[216] WARDHAUGH, R.

An introduction to sociolinguistics.

John Wiley & Sons, 2010.

[217] WEBSTER, S., DAVIDSON, J., BIFULCO, A., GOTTSCHALK, P., CARETTI, V., PHAM, T., GROVE-HILLS, J., TURLEY, C., TOMPKINS, C., CIULLA, S., ET AL.

European online grooming project (final report).

European Commission Safer Internet Plus Programme, Tech. Rep. (2012).

[218] WEIMANN, G.

Terrorism in cyberspace: The next generation.

Columbia University Press, 2015.

[219] WEISS, S. M.

Small sample error rate estimation for k-nn classifiers.

IEEE Transactions on Pattern Analysis and Machine Intelligence 13, 3 (1991), 285–289.

[220] WEISS, S. M., INDURKHYA, N., ZHANG, T., AND DAMERAU, F.

Text mining: predictive methods for analyzing unstructured information.

Springer Science & Business Media, 2010.

[221] WESTLAKE, B., BOUCHARD, M., AND FRANK, R.

Comparing methods for detecting child exploitation content online.

In *Intelligence and Security Informatics Conference (EISIC), 2012 European* (2012), IEEE, pp. 156–163.

[222] WESTLAKE, B. G., BOUCHARD, M., AND FRANK, R.

Finding the key players in online child exploitation networks.

Policy & Internet 3, 2 (2011), 1–32.

[223] WHITTY, M. T.

Anatomy of the online dating romance scam.

Security Journal 28, 4 (2015), 443–455.

[224] WINTERS, G. M., AND JEGLIC, E. L.

Stages of sexual grooming: Recognizing potentially predatory behaviors of child molesters.

Deviant Behavior 38, 6 (2017), 724–733.

[225] WOLFRAM, W. A.

A Sociolinguistic Description of Detroit Negro Speech. Urban Language Series, No. 5.

ERIC, 1969.

[226] YAN, X., AND YAN, L.

Gender classification of weblog authors.

In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs* (2006), pp. 228–230.

[227] YANG, S., KUROSE, J., AND LEVINE, B. N.

Disambiguation of residential wired and wireless access in a forensic setting.

In *INFOCOM, 2013 Proceedings IEEE* (2013), IEEE, pp. 360–364.

[228] YU, H., HO, C., JUAN, Y., AND LIN, C.

Libshorttext: A library for short-text classification and analysis.

Rapport interne, Department of Computer Science, National Taiwan University. Software available at http://www. csie. ntu. edu. tw/˜ cjlin/libshorttext (2013).

[229] YULE, C. U.

The statistical study of literary vocabulary.

Cambridge University Press, 2014.

[230] ZANASI, A.

Virtual weapons for real wars: Text mining for national security.

In *Proceedings of the International Workshop on Computational Intelligence in Security for Information Systems CISIS'08* (2009), Springer, pp. 53–60.

[231] ZELIKOVITZ, S., AND HIRSH, H.

Improving short text classification using unlabeled background knowledge to assess document similarity.

In *Proceedings of the seventeenth international conference on machine learning* (2000), vol. 2000, pp. 1183–1190.

[232] ZHANG, C., AND ZHANG, P.

Predicting gender from blog posts.

University of Massachussetts Amherst, USA (2010).

[233] ZHENG, R., LI, J., CHEN, H., AND HUANG, Z.

A framework for authorship identification of online messages: Writing-style features and classification techniques.

Journal of the Association for Information Science and Technology 57, 3 (2006), 378–393.

[234] ZIJLSTRA, H., VAN MEERVELD, T., VAN MIDDENDORP, H., PENNEBAKER, J. W., AND GEENEN, R.

De Nederlandse versie van de 'Linguistic Inquiry and Word Count' (LIWC).

Gedrag Gezond 32 (2004), 271–281.

[235] ZIPF, G.

Selected studies of the principle of relative frequency in language.

Harvard University Press, 1932.