A Likelihood Ratio Based Forensic Text Comparison with Multiple Types of Features

Master of General and Applied Linguistics (Advanced)

The Australian National University

Mr Sirawit Sodabanlu

November, 2021

This thesis is submitted in partial fulfillment of the requirements for the degree of Master of General and Applied Linguistics (Advanced) in the College of Arts and Social Sciences.

I hereby declare that, except where it is otherwise acknowledged in the text, this thesis represent my own original work.

All versions of the submitted thesis (regardless of submission type) are identical.

Sirawit Sodabanlu

November 2021

This thesis did not require human research ethics approval.

Table of Contents

Table of Figures

Table of Tables

Tables of Equations

Abstract

This study aims at further improving forensic text comparison (FTC) under the likelihood ratio (LR) framework. While the use of the LR framework to conclude the strength of evidence is well recognised in forensic science, studies on forensic text evidence within the LR framework are limited, and this study is an attempt of alleviating this situation. There have already been initiatives to obtain LRs for textual evidence by adopting various approaches and using different sets of stylometric features. (Carne & Ishihara, 2020; Ishihara, 2014, 2017a, 2017b, 2021). However, only few features have been tested in the similarity-only score-based approach (Ishihara, 2021), and there are many features left to be further investigated. To achieve the aim of the study, we will investigate some of the features in LR-based FTC and demonstrate how they contribute to the further improvement of the LR-based FTC system. Statistic, word n-gram (n=1,2,3), character n-gram (n=1,2,3,4), and part of speech (POS) n-gram (n=1,2,3) features were separately tested first in this study, and then the separately estimated LRs were fused for overall LRs. The databased used was prepared by Ishihara (2021), and the documents of comparison were modelled into feature vectors using a bag-of-words model. Two groups of documents, which both contained documents of 700, 1,400, and 2,100 words, were concatenated for each author, resulting in the total of 719 same-author comparisons and 516,242 different-author comparisons. The Cosine similarity was used to measure the similarity of texts, and the similarity-only score-based approach was used to estimate the LRs from the scores of similarity (Helper et al., 2012; Bolck et al., 2015). Log-likelihood ratio cost (C_{llr}) and their composites—C_{llr}^{min} and C_{llr}^{cal}—were used as assessment metrics. Findings indicate that (a) when the LRs of all the feature types are fused, the fused C_{llr} values are 0.56, 0.30, and 0.19 for 700, 1,400, and 2,100 words, respectively, and (b) feature selection depending on the nature of an FTC task matters to the performance of the FTC system and can contribute to the improvement of LR-based FTC.

Acknowledgement

This Master's thesis would not have been possible without the assistance of my supervisor, Dr. Shunichi Ishihara. Apart from your incredible passion and understanding towards forensic linguistics, you are one of the most thoughtful people I have ever known. You have not only assisted with academic support but also cared for me as a person for the past two years. I highly appreciate your efforts in making this thesis possible and will always look up to you as a role model.

I would like to express my gratitude towards Dr. Susy Macqueen, the convenor of the Master of General and Applied Linguistics at the Australian National University, whose support has been invaluable to me throughout my Master's degree. I would also like to give my thanks towards all of the academics at the ANU and Chiang Mai University for having made me who I am today.

My love and thoughts go towards my family and friends. You all are always there through my ups and downs, and no words can describe how grateful I am for your presence in my life. You all will always have a special place in my heart.

Chapter 1 Introduction

Since its introduction in the 1990s, the Internet has proliferated in human life in almost every corner of today's world. Despite being such a world-changing innovation, the Internet inevitably brings out the worst of people. They take advantage of the state-of-the art online platform to serve their unjust causes, including, but not limited to, identity fraud, data theft, privacy breach, and spam. One of the most popular vessels of these malicious intents is text. Text is highly accessible: most people learn to write through some way since they were young, and once it is on the Internet, it does not take much effort to comprehend it. Besides, text is one of the forms of communication that can spread quickly as devices of modern world do not require significant data or performance usage to read text.

It does not come as a surprise at all that people would want to put up some defensive measures to protect not only themselves but also society from text-based cybercrimes. Such urgent needs have been attracting attention to authorship analysis studies, or the studies that aim to answer the big question of authorship of texts in comparison. This thesis is about one type of authorship analysis: Forensic text comparison (FTC), the scientific evaluation of text to be submitted as evidence in court. In this chapter, I will discuss some background relevant to FTC and the framework that I will be using in this thesis, the likelihood ratio (LR) framework. Eventually, I will present my research aim and research questions, two navigation tools that will help us see through this thesis.

1.1 Forensic Text Comparison

As discussed earlier, FTC is the scientific evaluation of text to be submitted as evidence in court. It typically involves comparing two sets of text, usually of the offender (i.e. questioned) origin, and the suspect (i.e. known) origin, to assist the trier-of-fact in concluding if the suspect is guilty. FTC has its roots in authorship analysis. Authorship analysis is a scientific attempt to examine characteristics of text so as to conclude its authorship. Modern authorship analysis makes use of techniques from different disciplines—for example statistics, computer science, linguistics, and so on. Authorship analysis studies have been progressing through three major authorship analysis tasks, which are authorship identification/attribution, authorship verification, and authorship profiling (Rocha et al., 2016; Stamatatos, 2009). Authorship identification/attribution is the identification of the anonymous author of text, authorship verification concerns the verification of text whether it has been written by a certain author, and authorship profiling deals with the information on the characteristics—for example age, sex, education, nationality, etc.—of the author of text. In the case of FTC, authorship verification is thus considered the most similar process to FTC since they both look at the authorship of two different sources, the suspect/known and the offender/questioned texts. However, the main difference between authorship verification and FTC is that while the former aims to solve the problem, the latter aims to assist with the problem only (Ishihara, 2014, 2017a, 2017b). As has been mentioned earlier, FTC only compares texts of two sources—one from the questioned origin and another the known origin—and then computes the strength of textual evidence; thus, the term 'FTC', and not the other terms such as 'forensic text analysis', 'forensic text authorship analysis', or 'forensic text classification', is preferred here.

There are a number of studies that claimed themselves to be the studies that aimed to answer the big question of authorship the forensic context (Belsivi et al., 2020; de Vel et al., 2001; Grant 2007, 2010; Lambers & Veenman 2009; Zheng et al. 2003). That said, they treated the tasks at hand as classification problems. Therefore, those studies did not pay much attention to the textual data as evidence whose strength need to be quantified and submitted to court. FTC as is conducted in this thesis does not treat the task at hand as a classification problem, but it rather sees the task as an assistance to the court. In the court, the forensic expert should only provide the strength of evidence without expressing their opinion that may tamper with the court's belief or judgement. This will be explained in more detail in Section 1.2.

To conclude the strength of evidence and submit it to for the court's consideration, one needs a proper way of doing so. A likelihood ratio (LR) framework is widely in forensic identification science to conclude the strength of forensic evidence is the form of an LR (see Section 1.2). Some think of the LR framework as the only legally and logically correct framework for concluding the strength of evidence and thus advocate the use of the LR framework in forensic identification science (Aitken 1995; Aitken & Stoney 1991; Aitken & Taroni 2004; Balding & Steele 2015; Evett 1998; Robertson & Vignaux 1995), while some think that the LR framework is one of several available forensic frameworks that are capable of presenting the strength of evidence (Lund & Iyer, 2017). In LR-based FTC, the forensic expert is expected to provide the trier-of-fact with the strength of text evidence in the form of an LR. Having considered all pieces of evidence including the textual evidence, the trier-of-fact is responsible for making a final decision regarding the case, i.e. guilty or not guilty, by taking into account their former belief regarding the case which was made by the other evidence relating to the case. Despite establishing a logically and legally correct tool for analysing text evidence and the presence in forensic science studies (see Section 1.3), FTC under the application of LR framework has been considered lagging behind the other branches of forensic science (Ishihara, 2014, 2017a, 2017b). Consequently, LR-based FTC needs not only academic but also professional attention for it to serve justice for society.

1.2 The Likelihood Ratio Framework and Bayes' Theorem

It is widely discussed in the forensic community that the role of the forensic expert in the court is only to provide the trier-of-fact with the strength of evidence relevant to the case, and not to make an implication for the suspect's presumption of innocence in any way (Aitken, 1995; Aitken & Stoney, 1991; Aitken & Taroni, 2004; Balding & Steele, 2015;

Evett 1998; Robertson & Vignaux, 1995). In doing so, the LR framework is employed to quantify the strength of evidence. This can be seen as the ratio of the probability that the evidence (*E*) would occur if a hypothesis (H_p) is true and the probability that the same evidence would occur if the alternative hypothesis (H_d) is true. In the legal context, H_p and H_d (1) are seen as ‘prosecution’ and ‘defence’ hypotheses respectively, and in FTC, H_p and H_d are referred to as ‘same-author’ and ‘different-author’ hypotheses respectively. The LR equation is formulated in Equation (1) below:

$$\text{LR} = \frac{p(E|Hp)}{p(E|Hd)} \qquad ($$

Equation (1) shows the LR as the ratio of two conditional probabilities: the probability of *E* given H_p and the probability of given H_d. To put it simply, the LR shows the likelihood that *E* would be found in the H_p scenario (p($E|H_p$)) against the likelihood that the same *E* would occur in the H_d scenario (p($E|H_d$)). It is widely discussed in the forensic community that it is only p($E|H$) (i.e. the likelihood of *E* given a hypothesis), and not p($H|E$) (i.e. the likelihood of a hypothesis given *E*), that is considered the province of the province of forensic expert; therefore, the forensic expert may only assess the strength of *E*, and never that of H_p or H_d. (Aitken, 1995; Aitken & Stoney, 1991; Aitken & Taroni; 2004; Hicks et al., 2015). If the numerator p($E|H_p$) is higher than the denominator p($E|H_d$), the LR will be higher than one, meaning that it is more likely that the evidence is more likely to occur under the H_p scenario rather than under the H_d scenario. With respect to FTC, this means that the evidence to be observed if the textual evidence is written by the same author rather than written by different authors. Meanwhile, if the numerator p($E|H_p$) is lower than the denominator p($E|H_d$), the LR will be lower than one, meaning that the evidence is more likely to occur under the the H_d scenario rather than the the H_p scenario. This means that for FTC cases, the evidence is more likely to be observed if it is written by different authors rather than written by the same author. From the explanations on the LR equation above, it can be concluded that the LR higher than one supports the H_p (the same-author hypothesis), while the LR value lower than one supports the H_d (the different-author hypothesis). That said, the LR is not a binary expression of whether either hypothesis is true. The LR is instead a gradient encapsulation of how far the LR is from unity (i.e. one) and therefore how strong the evidence is for, or against, either hypothesis. As far as it moves away for one, the LR provides greater support for either H_p (in case that LR>1) or H_d (in case that LR<1) than LRs that are closer to one. On the other hand, as close as it sticks to the unity threshold, the LR provides less support for either H_p (in case that LR<1) or H_d (in case that LR>1) than LRs that are further from one. It is worth pointing out again that the LR only implies the strength of evidence, not the strength of hypothesis. This means that the LR can tell only the evidence is more likely to arise under one hypothesis than the other, and not that the evidence is written by the same author than different authors, or vice versa. For instance, an LR of 100 would mean that it is 100 times more likely for the evidence to be observed if the evidence is written by the same author rather than by different authors. That said, an LR of 100 would not mean that it is 100 times more likely that the evidence is written by the same author rather than by different authors.

Although the strength of hypothesis is not considered the province of the forensic expert and they does not know the former belief regarding the case and the strength of other pieces of evidence, the forensic expert may use Bayes' Theorem to calculate the strength of hypothesis. Bayes' Theorem in the odds form is provided in Equation (2) as follows:

$$\underset{\text{Posterior odds}}{\frac{p(Hp|E)}{p(Hd|E)}} = \underset{\text{Prior odds}}{\frac{p(Hp)}{p(Hd)}} \text{ x } \underset{\text{Strength of evidence (LR)}}{\frac{p(E|Hp)}{p(E|Hd)}} \qquad (2)$$

Equation (2) illustrates how Bayes' Theorem can be employed to calculate the strength of hypothesis, or what is called the posterior odds. To describe, the posterior odds is the result of the trier-of-fact's beliefs of two competing hypotheses relevant to the case, or what is called the prior odds, being updated by the strength of evidence, which is expressed through the LR.

In legal casework, more than one piece of evidence related to the case at hand is often expected; therefore, the final LR may be composed of the strength of several pieces of evidence (see Section 2.5.3). After the strength of all pieces of relevant evidence is presented to the trier-of-fact, the trier-of-fact will take into account how that reinforces or refutes the competing hypotheses or their belief, which is eventually reflected in the strength of hypothesis or the so-called posterior odds.

In legal casework, the forensic expert should and must not try to usurp the role of the trier-of-fact by influencing or implying the posterior odds or the prior odds in any way possible; otherwise, there would be serious logical and legal consequences (Aitken 1995; Evett, 1998). Logically speaking, in order to calculate the posterior odds, the forensic expert needs to also know the prior odds, which is only privy to the trier-of-fact. Moreover, in order to make an informed judgement towards the case, the trier-of-fact needs to take into account not only single piece of evidence (i.e. textual evidence) but every piece of evidence related to the case. In terms of legal wrongdoing, the forensic expert's involvement in calculating the posterior odds, although logically they cannot, also violates the law as the final decision of whether someone is guilty or innocent should solely be the trier-of-fact's responsibility. This is seen as usurping the judicial system's role and is considered illegal. In short, the forensic expert cannot logically assess the strength of hypothesis since they do not have access to other important pieces of information (i.e. prior odds and LRs of other evidence), and they cannot legally do so as it is illegal to try overreaching the role of the court. The province of the forensic expert should therefore be to assess the strength of evidence only, and nothing more in the court.

1.3 Forensic Text Comparison within the Likelihood Ratio Paradigm

There are a handful of LR-based FTC studies that I will use to reference for my thesis as these are the extended versions of the other LR-based FTC studies (Carne & Ishihara, 2020; Ishihara, 2014, 2017a, 2017b, 2021). Despite these valuable initiatives, there are still several areas left to be investigated in LR-based FTC.

One of those areas is the stylometric features untested in LR-based FTC, the main subject matter of this thesis. In LR-based FTC, only a handful of features have been tested, for example word- and character-based statistics (Carne and Ishihara, 2020; Ishihara, 2017a, 2017b, 2021), word n-gram features (Ishihara, 2017b), and character n-gram features (Ishihara, 2014, 2017b). Moreover, most of the tested features have been tested in LR-based FTC studies that adopted the feature-based and the similarity-typicality score-based approaches, both of which are not what this thesis had adopted (see Section 2.5.1). This thesis aims to test the efficiency of the features both untested and tested in LR-based FTC within the similarity-only score-based approach. For more information on the features tested in this thesis, see Section 2.4.1 and Section 2.4.2.

In authorship analysis tasks, what determines the performance of features is the number of features used for each feature set or experiment setting. In LR-based FTC studies, the optimal number of features that yields the best FTC system performance depends on the nature and the methodology of an FTC task. For instance, in Ishihara (2021) that experimented FTC tasks on an authorship verification corpus and adopted the similarity-only score-based approach, the number of the most frequent words that yielded the best results is 260. Meanwhile, in Carne and Ishihara (2020) that worked on the same corpus but used the different approach (i.e. the feature-based approach), the optimal number of the most frequent words is 180. As an LR-based FTC study within the similarity-only score-based approach, this thesis aims to empirically test how many features within a specific feature type that the FTC system specifically designed for this thesis needs to obtain optimal results and discuss the underlying reasons for such a phenomenon.

Also worth investigating is whether the performance of the FTC system can be improved by means of calibration. In LR-based FTC studies, logistic regression calibration (LRC; Brümmer & du Preez, 2006) is employed due to its being a robust technique that is commonly applied in LR-based forensic studies. The LRC is used to optimise the quality of the LRs that have been converted from scores of similarity and may not yet be well calibrated. The effects of the LRC differ depending on the approach for estimating LRs adopted. For example, the LR-based FTC studies within the similarity-typicality score-based approach and the feature-based approach, demonstrated that logistic regression calibration was needed to optimise the quality of the LRs derived (Carne & Ishihara, 2020; Ishihara, 2014, 2017a, 2017b). However, Ishihara (2021), the LR-based FTC study within the similarity-only score-based approach, demonstrated that the LRC was not needed. As an LR-based FTC study within the similarity-only score-based approach, this thesis aims to empirically test whether the LRC was needed for the FTC system specifically designed for this thesis and the reasoning why it is so.

1.4 Research Aim and Research Questions

As per the discussed background and literature relevant to this thesis, LR-based FTC is still in its infancy as it has yet to be recognised in the wider forensic science community. A number of areas are left to be investigated, and many contributions can be made to LR-based FTC. Therefore, the aim of this thesis is to further improve LR-based FTC by investigating some selected areas—untested stylometric features, the optimal number of features, and further optimisation of the FTC system—as mentioned in Section 1.3. To achieve the research aim, I will posit the research questions as followed:

RQ1 How does the inclusion of a specific feature within a feature type improve or deteriorate the performance of the LR-based FTC system?

RQ2 How does the inclusion of a specific feature type improve or deteriorate the performance of the LR-based FTC system?

RQ3 For the database used in this thesis, what is the optimal number of features for each feature set in the FTC system designed for this thesis?

RQ4 Does the FTC system need further optimization, namely logistic regression calibration?

Chapter 2 Methodology

In typical FTC research, the procedures are usually broken down into the following stages for estimating the LRs: database, database partition, tokenisation, feature extraction, and testing.

2.1 Database

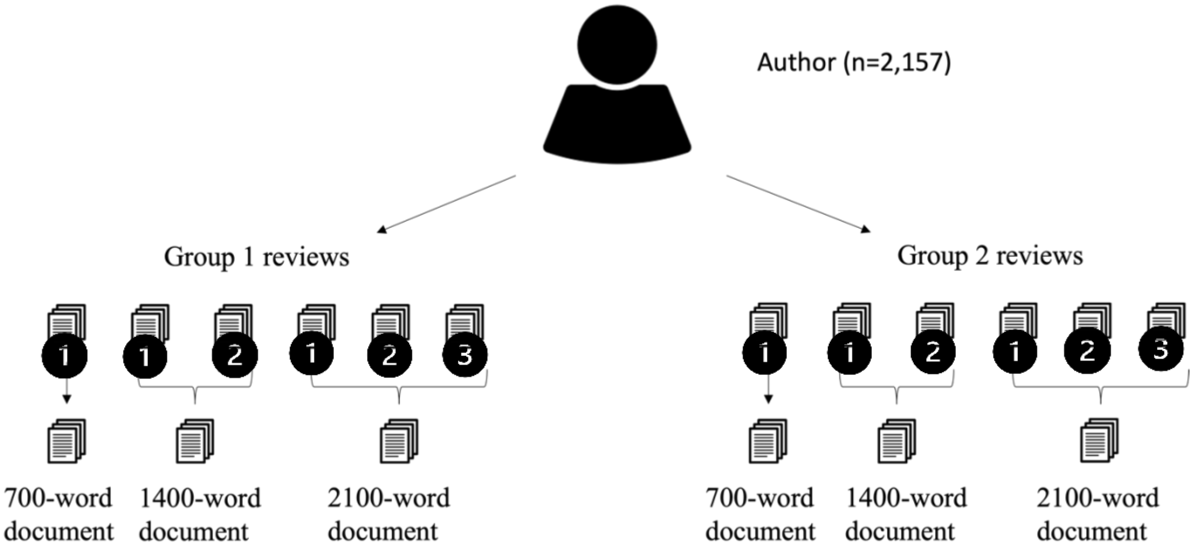
This thesis makes use of the same database as was used in Carne and Ishihara (2020) and Ishihara (2021), and some methodological details in this chapter would be adapted from Ishihara (2021). The following is the summary of the database used in this thesis.

The database used was made by compiling data from the Amazon Product Data Authorship Verification Corpus (Halvani et al., 2017), which is again based on the Amazon Product Data Corpus (He & McAuley, 2016). Ishihara (2021) selected only the authors who wrote six or more product reviews in which the word length exceeds 700 words, were singled out, resulting in a total amount of 2,157 authors. Only first six reviews of each author would be used. The six reviews were then separated into two groups: the first three reviews and the last three reviews.

Three different documents that differed in word length (700, 1,400, and 2,100) were to be created using these two groups of reviews for each author by means of concatenation. Word length was controlled for a reason: the word lengths of 700, 1,400, and 2,100 words allow us to see how the FTC system performance fluctuates with word length. Figure 1 visually demonstrates how the two groups' reviews were used to generate three documents of different word lengths. According to Figure 1, the first review of each group was used as is to create a 700-word document for each group. Then, for each group, the first and second reviews were concatenated into a document of 1,400 words, and the first, second, and third reviews into a document of 2,100 words. This means that each author (n=2,157) had

Figure 1

Concatenation of Three Review Texts for Generating Documents of Different Word Lengths for Each Group of Documents for Each Author (n=2,157)



Note. The number '1', '2', or '3' suggests that the document attached is the first, second, or third document respectively of either group, out of two groups, of one author.

two separate sets of documents which each contained three documents of different word lengths (700, 1,400, and 2,100).

The use of this database deserves some justification. Ishihara (2021) saw the use of this database to conduct an FTC experiment as the most suitable simulation of 'the forensic scenario of one-to-many communication'. Currently, there are no available databases that contain forensic text evidence or that can be used specifically for FTC experiments. Although it was specifically designed for authorship verification tasks and not for forensic purposes, this database seemed to be the most appropriate database. Not only the ample amount of data is concerned, but also the nature of product reviews themselves. Ishihara (2021) claimed that the content in product reviews aims to 'convey one's views to others', which usually resembles that of criminal texts (e.g. ransom notes, death threats, defamation on social media, etc.). That said, they usually differ in that product reviews tend to be written in a more neutral tone while criminal texts in a more aggressive, dramatic tone.

2.2 Database Generation

The database would further be divided into the following three sub-databases: test, background, and development. The visual representation of the sub-database generation is displayed in Figure 2. In Figure 2, all the three sub-databases would each contain 719 authors and their associated documents of different word lengths, and each author would have two groups of documents which differed in word length.

2.2.1 Test Database

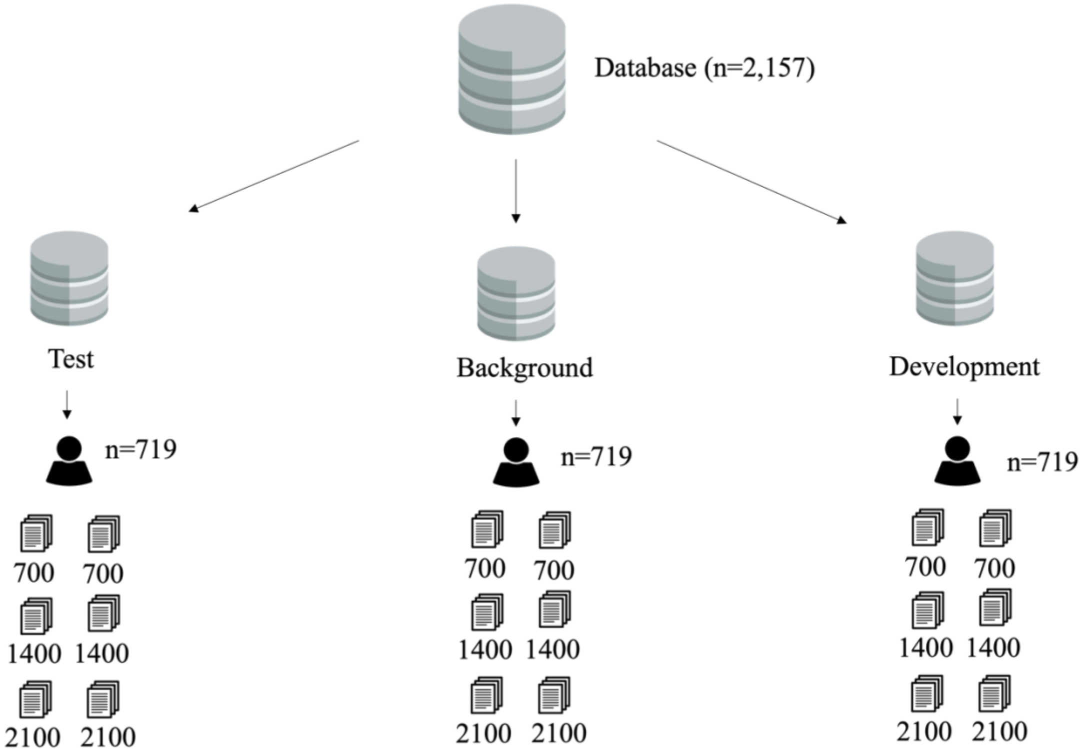
The test database was used for assessing the performance of the FTC system. To explain, the document stored in the test database would be used for simulating same-author and different-author comparisons. This means that 719 authors of the test database would result in 719 same-author comparisons and 516,242 different-author comparisons. While the generation of same-author comparisons is straightforward (i.e comparisons between two groups of documents per author), that of different-author comparisons is not and deserves some more explanation. The generation of different-author comparisons singled out two authors at a time to compare their documents, resulting in 258,121 (=$_{719}C_2$) different-author pairs of comparison. Since each author has two groups of documents of different word length, two independent comparisons between two authors were possible. This resulted in 516,242 (=$_{719}C_2$ x 2) different-author comparisons in total. These same-author and different-author comparisons were separately tested for each of the different word lengths.

2.2.2 Background Database

The background database served to train the score-to-LR conversion model (see Section 2.5.2). In LR-based FTC within the similarity-only score-based approach, the measured features of two text samples are compared to each other’s, and the forensic expert

Figure 2

Generation of Three Sub-databases of Test, Background, and Development



has to quantify the degree of similarity or difference between two text samples in the form of a score. However, the score obtained in this process is a simple measure of similarity between two documents; this is done by the Cosine distance (see Section 2.5.2.1). The forensic expert needs to convert the score to an LR using the score-to-LR conversion model trained by the same-author and different-author scores obtained from the background database. The background database also contains 719 authors, which also lead to 719 same-author comparisons and 516,242 different-author comparisons. This means that it would be expected to obtain 719 same-author scores and 516,242 different-author scores which would be used to train the score-to-LR conversion model later. The mapping from scores to LRs is regarded as essential to LR-based FTC within the similarity-only score-based approach since only the well-calibrated LRs, and not the scores, can be interpreted as the strength of evidence.

2.2.3 Development Database

Despite having converted scores to LRs by using the score-to-LR conversion model, the forensic expert may find the LRs obtained cannot be interpretable as the strength of evidence; this may occur because of inappropriate modelling assumptions or the shortage of training data for the score-to-LR conversion model. Consequently, these uncalibrated LRs need to be calibrated through an appropriate calibration model. This calibration of uncalibrated LRs may be termed the second calibration; however, unlike the first calibration, it may not be obligatory. If the second calibration, causes no improvements to the FTC system performance—which means that the first calibration has already yielded well calibrated LRs—the second calibration may be deemed unnecessary. As part of the research questions, this thesis aims to investigate whether the second calibration is actually needed for LR-based FTC under the similarity-only score-based approach.

The calibration model used for this thesis is a logistic regression model, one that is widely used in LR-based forensic identification science (Morrison, 2013). At this stage, the LRs derived from the same-author and different-author scores obtained from the development database were used to train the logistic regression model, which would be used to calibrate the LRs derived from the score-to-LR conversion model. For more information on logistic regression calibration, see Section 2.5.3.

2.3 Tokenisation

Tokenisation is the process of segmenting text into smaller linguistic units. For the purposes of this thesis, the text data were tokenised using the ‘quanteda’ R statistical package (Benoit et al., 2018). As part of the ‘quanteda’ library, The ‘token()’ function was utilised. Stemming and lemmatization were done were not applied; this means that ‘review’, ‘reviews’, and ‘reviewing’ would be treated as different words. Nothing had been done to punctuations and special characters, so all of these would be treated as separate word tokens from the words to which they were attached.

2.4 Feature Extraction

Computer-based authorship algorithms cannot handle raw text; text data need to be properly represented according to authorship features. This is the definition of feature extraction, or the process of reducing raw, messy data into something that is more manageable and easier for the computer to process. For the databased used, feature extraction was made possible with the use of a bag-of-words approach.

In text processing, a bag-of-words approach is a text modelling technique that is used widely. Within a bag-of-words approach, text data are modelled into a feature vector, which subsequently contains the feature values which are calculated for the text. Despite the so-called name, a bag-of-words approach does not only handle words as other types of items, for example characters and part-of-speech tags, can also be handled by such an approach. Despite being praised for its simplicity and effectiveness (Diederich et al., 2003), a bag-of-word approach has one major downfall: it does not take into account the sequential nature of linguistic items appearing in written text. This is where an n-gram approach comes in and shines. An n-gram is a contiguous string of every n linguistic item(s) appearing in the text which, when being modelled, allows the text’s sequential information to be intact. Some even consider an n-gram approach to text modelling as an extension of a bag-words-approach (Mutinda et al., 2021; Stefanovič et al., 2019). This is because an n-gram approach still represents text via a spatial vector, but instead of single items represented via a bag-of-words approach, sequences of items (e.g. sequences of words, characters, or part-of-speech tags) with their relative frequencies are used as feature values; therefore, sometimes an n-gram approach is even called a bag-of-n-grams approach (Stefanovič et al., 2019).

FTC operates according to one fundamental principle: people have their own styles of writing (McMenamin, 2001, 2002). Writing styles can be captured through a collection of quantifiable units in text which are called stylometric features. While Rudman (1997) posited there were thousands stylometric features that can be used to quantify authorship, only a handful of them have previously been tested in LR-based FTC (Carne & Ishihara, 2020; Ishihara, 2014, 2017a, 2017b, 2021).

Basic statistics of both words and characters and word, character, and part-of-speech n-grams, were chosen as stylometric features in this thesis. Using these stylometric features, each document or text would be modelled via a bag-of-words model. Each type of stylometric features will be discussed in detail.

2.4.1 Statistical Features

The first type of stylometric features tested in the FTC system is statistical features (henceforth SFs). In this thesis, SF are features that are calculated using statistical measures, such as the mean and the standard deviation, of the raw data. In authorship analysis tasks, these features are either word- or character-based, so they are subsequently categorised as lexical or character features respectively (Zheng et al., 2006; Stamatatos, 2009; Rocha et al., 2017). However, such features are termed statistical features for this thesis for data modelling and presentation reasons. There are ten SFs tested in this thesis. Table 1 displays all the ten SFs with brief descriptions.

The inclusion of the ten SFs is based on previous research on authorship analysis. Previous authorship studies including Iqbal et al (2010), de Vel et al. (2001), Grieve (2007), and Zheng et al. (2006) demonstrated that the stylometric features displayed in Table 1 yield good results and therefore are robust for authorship analysis tasks. There are some initiatives aimed at testing these stylometric features within the LR framework. Within the LR framework, Ishihara (2017a, 2017b) demonstrated that most of the statistical features in

Table 1

List of 10 Statistical Features Tested in the Statical Feature Experiment

Statistical Features	Descriptions
1. TTR	Type-token ratio (the total number of unique words (types) divided by the number of words (tokens) in text)
2. K	Vocabulary richness as defined by Yule
3. Hapax	Hapax legomenon (i.e. words that appear only once)
4. FK	Flesch-Kincald readability score (i.e. how difficult it is to understand one text in English)
5. DigitRatio	The ratio of digits
6. PuncRatio	The ratio of punctuation marks (eight punctuation marks , . ? ! ; : ' \ ")
7. SpecialCharRatio	The ratio of special characters (twenty two special characters < > % | [ ] } { @ # ~ + - * $ ^ & = \ / ( ))
8. UppercaseRatio	The ratio of uppercase characters
9. CharNumberPerWord	The average number of characters per word
10.FreqUnusualWordUS	The frequency of unusual words (en_US dictionary)

Table 1 are robust for LR-based FTC tasks. Some of the features, for example hapax legomenon and Flesch-Kincald readability scores, have never been tested in research on the LR-based FTC, so this thesis will be the first of its kind to test these two features within the LR paradigm. The usefulness of hapax legomenon has been explored by Savoy (2012a, 2012b) in authorship attribution tasks; Savoy found that it can contribute to a good classification algorithm when used with other features that also focus on extracting information on specific vocabulary. Regarding the Flesch-Kincaid readability scores, to the best of my knowledge, this feature has not seen to be tested in authorship analysis studies. However, the Flesch-Kincaid readability gets extensively used in clinical research (Eloy et al., 2012; Johnstone & Giles, 2017) to analyse the readability of clinical materials, so I reckoned this feature may be a discriminating feature when it came to analysing the readability of product reviews and therefore the authors behind them.

Among to the ten SFs tested, there are five that deserve more explanations, which are the number 1, type-token ratio (TTR), the number 2, Yule's K (K), the number 3, hapax legomenon (Hapax), the number 4, Flesch-Kincald readability score (FK), and the number 10, the frequency of unusual words (FreqUnusualWordUS). TTR, K, and FK were computed using the package 'quanteda' in R (Benoit et al., 2018), while FreqUnusualWordUS was computed using the package 'hunspell' in R (Ooms, 2020).

TTR stands for type-token ratio, which is the total number of unique words (types) divided by the total number of words (tokens) in text. The formula used to calculate TTR is displayed as in Equation (3):

$$TTR = \frac{V}{N} \tag{3}$$

where *V* refers to the total number of types and *N* refers to the total number of tokens.

K is the measurement of vocabulary richness proposed by George Udny Yule (1944), a British statistician. The formula used to calculate Yule's K for this experiment is displayed as in Equation (4):

$$K{=}10^4{\times}[-\frac{1}{N} + \sum_{i=1}^{v} F_v(i{,}N)(\frac{i}{N})^2] \tag{4}$$

where *V* refers to the number of types, *N* refers to the number of tokens, and $F_v(i,N)$ refers to the number to types occurring *i* times in the length *N* (Yule, 1944, as cited in Tweedie & Baayen, 1988, p. 330).

Two vocabulary richness features, namely TTR and Yule's K, were selected for this thesis since for different reasons. As TTR is merely the ratio between the total number of word types and that of word tokens in disregard of the length of the whole test, it is not robust to be used with texts of different length. Therefore, another vocabulary richness measure that considers the weight of each word type given the total number of word tokens, which in this case is Yule’s K, is used in this thesis to help deal with texts of different length.

FK, or Flesch-Kinclad readability score, is the measurement used to calculate the readability of text, which was developed by Rudolf Flesch and J. Peter Kincald (Flesch, 1948; Kincald et al., 1975). The formula used to calculate Flesch-Kinclad readability score is displayed as in Equation (5):

$$\text{FK} = 0.39\times ASL+11.8\times\frac{N_{sy}}{N_w}-15.59 \tag{5}$$

where *ASL* refers to the average length of the sentence (the number of words divided by that of sentences), N_{sy} refers to the number of syllables, and N_w refers to the number of words.

Hapax is shortened for the frequency of hapax legomena, words that appear only once in text. The formula used to calculate Hapax is displayed as Equation (6):

$$\text{Hapax} = \frac{Hap}{N_t} \tag{6}$$

where *Hap* refers to the number of hapax legomena, and N_t refers to the number of tokens in text.

FreqUnusualWordUs is the frequency of unusual words based on the American English dictionary called ‘en_US dictionary’. As part of the ‘hunspell’ R package, the en_US dictionary serves as a baseline for detecting misspelt or unusual words that do not appear in the dictionary. The en_US is an open spelling dictionary that is commonly used in natural(7) language processing tasks. The formula used to calculate FreqUnusualWordUs is displayed as in Equation (7):

$$\text{FreqUnusualWordUs} = \frac{N_{uw}}{N_t}$$

where N_{uw} refers to the number of words that do not appear in the en_US dictionary, while N_t refers to the number of tokens in text.

2.4.2 N-gram Features

Another type of stylometric features that would be tested for this thesis is n-gram features. Three sub-types of n-gram features were tested, which are word n-grams (WNGs), character n-grams (CNGs), and part-of-speech n-grams (PNGs).

2.4.2.1 Word N-gram Features. A word n-gram (WNG) is a continuous sequence of every n word(s) appearing in the text. Previous authorship analysis literature, for example Coyotl-Morales et al. (2006) and Sanderson and Guenter (2006), had proposed the use of word n-gram features to acquire word-order and contextual information of text. This thesis would be using WNGs with an n of one (word unigrams; WN1s), two (word bigrams; WN2s), and three (word trigrams; WN3s). The number of features fixed for WNGs ranged from 5 features to 600 features at one experiment setting. That is, 5 to 600 sequences of n words would be tested separately for each word length. In LR-based FTC, Ishihara (2014, 2017b) used WNGs (n=1,2,3) in the FTC system and achieved good results with an ample amount of data. However, the technique to estimate the LRs in Ishihara (2014, 2017b) is different from the other one I adopted for this thesis (namely the similarity-only score-based approach; see Section 2.5.1); therefore, this thesis will be the first to trial WN1s, WN2s, and WN3s with application to LR-based FTC within the similarity-only score-based approach. This thesis made use of the token() function in the 'quanteda' R library (Benoit et al., 2018) to parse text into tokens and the tokens_ngram() function to construct WNGs.

2.4.2.2 Character N-gram Features. A character n-gram (CNG) is a continuous sequence of every character(s) appearing in the text. CNGs' efficiency in capturing contextual and stylistic information of text has been emphasised in a number of authorship analysis studies (Koppel et al., 2011; Stamatatos, 2013). This may partly result from that CNGs are easier to process than other types of features, making generated CNGs are supported by data more than being calculated from elsewhere. The number of CNGs tested at each experiment setting varies between the different CNGs; CN1s were tested only from 5 to 90 features, while CN2s, CN3s, and CN4s were tested from 5 to 1,000 features. The reason is that there are far less characters than words, resulting in the much smaller range of CN1s (5-90) than that of WN1s (5-600). Note that 90 CN1s include both uppercase and lowercase characters, special characters, and punctuation marks. For the testing range of CN2s and CN3s (5-1,000), as mentioned earlier, CNGs are considered more statistically auspicious than the other types of features as sequences of characters appear more often than sequences of words or part-of-speech tags. As with WNGs, CNGs have already been tested in LR-based FTC and yielded satisfactory results (Ishihara, 2014, 2017b). That said, this thesis will be the first in trialing using the similarity-only score-based approach to estimate the LRs derived with CN1s, CN2s, CN3s, and CN4s. Much like the case with WNGs, this thesis used the token() function of the same library to parse text into characters and the tokens_ngram() function to construct CNGs.

2.4.2.3 Part-of-speech N-gram Features. A part-of-speech n-gram (PNG) is a continuous sequence of every part-of-speech tag(s) appearing in the text. PNGs can be built in the same way as are word and character n-gram features, but it needs a robust part-of-speech tagger to accurately designate what part-of-speech of the words in the text is and thus to contribute to the usefulness of part-of-speech information in modelling authorial analysis. A handful of authorship analysis studies, including Diederich et al. (2003) and Kukushkina et al. (2001), have tested PNGs in their studies, and although they showed some promising signs, the results obtained from PNGs were not as satisfactory as those obtained from lower-level features such as WNGs or CNGs. Nonetheless, it is reported that PNGs, if modelled and trained correctly, can capture both morphological and syntactic information, something that WNGs and CNGs find hard to do (Gamon, 2004; Sidorov et al., 2014). For example, Sidorov et al. (2014) made use of PNGs in their authorship attribution task on a corpus of three authors as a baseline feature set and found that PNGs, especially PN2s and PN3s, generally perform better than WNGs and SNGs.

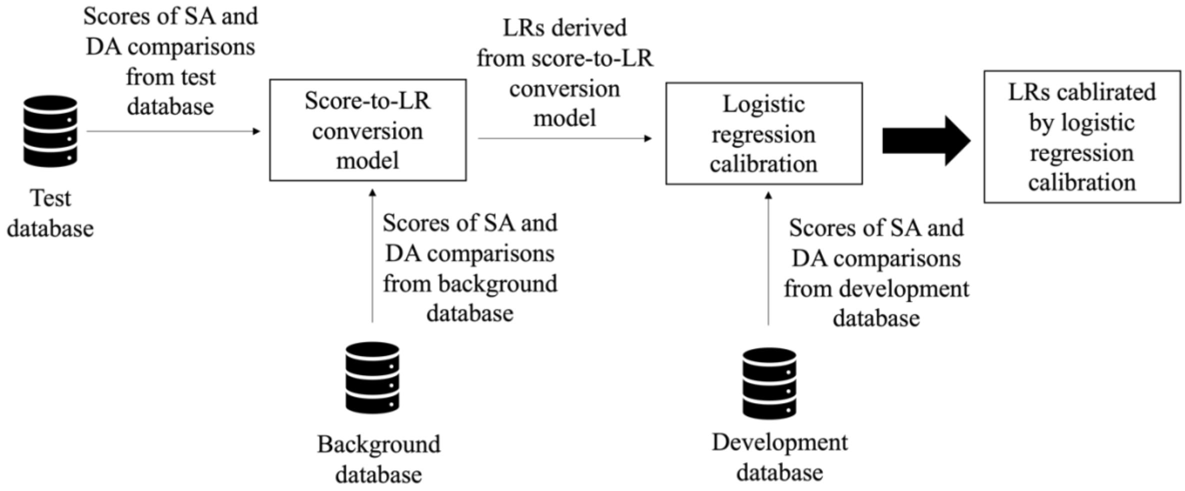
In this thesis, PN1s would be tested from 5 to 40 features, while PN2s and PN3s from 5 to 600 features. Part-of-speech tags are far less than words or characters, leading to that there were 40 part-of-speech tags that would be tested at one experiment setting for PN1s. PNGs have not yet been tested in LR-based FTC. This thesis will be the first of its kind to prove the efficiency of part-of-speech n-gram features in LR-based FTC within the similarity-only score-based approach. The 'spacy_parse()' function of the 'spacyr' R library (Benoit & Matsuo, 2020) was used to parse text of which the outcomes contain part-of-speech information. Using the part-of-speech tags, PNGs were constructed using the 'tokens_ngrams()' function of the 'quanteda' R library (Benoit et al., 2018).

2.5 Simulation of Same-author and Different-author Comparisons

The processes of the similarity-only score-based approach to calculating the LRs for the same-author and different-author comparisons of the test database are visualised in Figure 3. To explain, the simulation of the same-author and different-author comparisons was computed, and the outcomes were in the forms of scores. That said, these scores were not yet ready to be interpretable; they had to go through the score-to-LR conversion model, which was trained by the scores of the same-author and different-author comparisons from the background database. These LRs might or might be not well calibrated due to inappropriate modeling assumption or the shortage of data to train the score-to-LR conversion model. Therefore, the LRs derived from the score-to-LR conversion model might go through another calibration, which is logistic regression calibration (LRC). The LRC model was trained by the scores of the same-author and different-author comparison from the development database. Whether the FTC system in this thesis needs the LRC would be be empirically tested in this thesis.

Figure 3

Processes of Similarity-only Score-based Approach to Obtain LRs


Scores of SA and DA comparisons from test database
Test database
Score-to-LR conversion model
LRs derived from score-to-LR conversion model
Logistic regression calibration
LRs cablirated by logistic regression calibration
Scores of SA and DA comparisons from background database
Background database
Scores of SA and DA comparisons from development database
Development database

Note. 'SA' stands for 'same-author', while DA 'different-author'. Note that logistic regression calibration may or may not be needed depending on whether the LRs derived from the score-to-LR conversion model are well calibrated; this will be empirically tested in this thesis.

The metrics of performance used to assess the quality of the LRs derived are log-likelihood-ratio cost (*Cllr*), which can be broken down into a discrimination loss ($Cllr^{min}$) and a calibration loss ($Cllr^{cal}$). The magnitude of the LRs derived are to be visualised via Tippett plots.

2.5.1 Similarity-only Score-based approach

In forensic identification science, there are several approaches that the forensic expert can adopt to obtain the LRs for their evidence. For example, score-based approaches, either similarity-only or similarity-typicality, are widely used to estimate the LRs (Block et al., 2015; Chen et al., 2018; Helper et al., 2012; Leegwater et al., 2017). There is also an alternative approach to estimating the LRs, which is the featured-based approach (Aitken & Gold, 2013; Carne & Ishihara; 2020).

This thesis has made use of the similarity-scored based approach. Since there are several available approaches for estimating the LRs, the use of the similarity-only score-based approach warrants some justification. In calculating LRs, the similarity-only score-based approach considers only the similarity of the evidence and does not take into account the typicality of evidence. This results in that some forensic scientists did not agree with the use of the similarity-only score-based approach as they saw that both similarity and typicality are necessary to calculate LRs that can be used as the strength of evidence in the court (Morrison & Enzinger, 2018; Neumann & Ausdemore, 2020). Some studies (Block et al., 2015; Garton et al., 2020) also found that the magnitude of the LRs derived with the similarity-only score-based approach was likely to be smaller than those derived with the other approaches.

However, despite disadvantages, the similarity-only score-based approach is regarded as the approach that can handle the multidimensionality of features well (Bolck et al., 2015; Garton et al., 2020). The similarity-only score-based method converts the multidimensional feature space to a univariate score space. One advantage of the univariate score space is that it can be effectively trained and modeled by even a limited amount of data. Despite being weaker than derived with the other approaches, the magnitude of LRs derived with the similarity-only score-based approach was found it was more stable than derived with the other approaches (Block et al., 2015; Garton et al., 2020). Actually, Garton et al. (2020) argued for the similarity-only score-based approach that it is the only viable approach to obtain the LRs for the high dimensional evidence. Ishihara (2021) argued for the use of the similarity-only score-based approach that since the authorship features tested in FTC are highly dimensional and most of the time are correlated, the similarity-only score-based approach seems to be the most viable option in LR-based FTC. These facts regarding stylometric features pose some challenges to the other approaches of estimating the LRs for the evidence since they cannot effectively handle the multivariate structure of stylometric features and cannot consider the correlations between them.

2.5.2 Mapping from Score to Likelihood Ratio

In the similarity-only score-based approach, the scores of similarity obtained from the same-author and different-author comparisons from the test database, need to go through the score-to-LR conversion model. This results in the scores of similatiy being transformed into the LRs that, if well-calibrated, are ready to be interpreted as the strength of evidence. For this thesis, the Cosine similarity was used to calculate the scores of similarity between the documents of comparison (see Section 2.5.2.1). The LR of the score needed to be assessed against the probabilistic distributions trained from the same-source and different scores. The LR can then be mathematically expressed as in the following Equation (8):

$$\begin{aligned} \text{LR} &= \frac{\mathrm{f}(\Delta(\mathrm{x,y})|H_p)}{\mathrm{f}(\Delta(\mathrm{x,y})|H_d)} \\ &= \frac{\mathrm{f}(\Delta(\{w_1^x, w_2^x \ldots w_n^x\},\{w_1^y, w_2^y \ldots w_n^y)\}|H_p)}{\mathrm{f}(\Delta\{w_1^x, w_2^x \ldots w_n^x\},\{w_1^y, w_2^y \ldots w_n^y)\}|H_d)} \end{aligned} \tag{8}$$

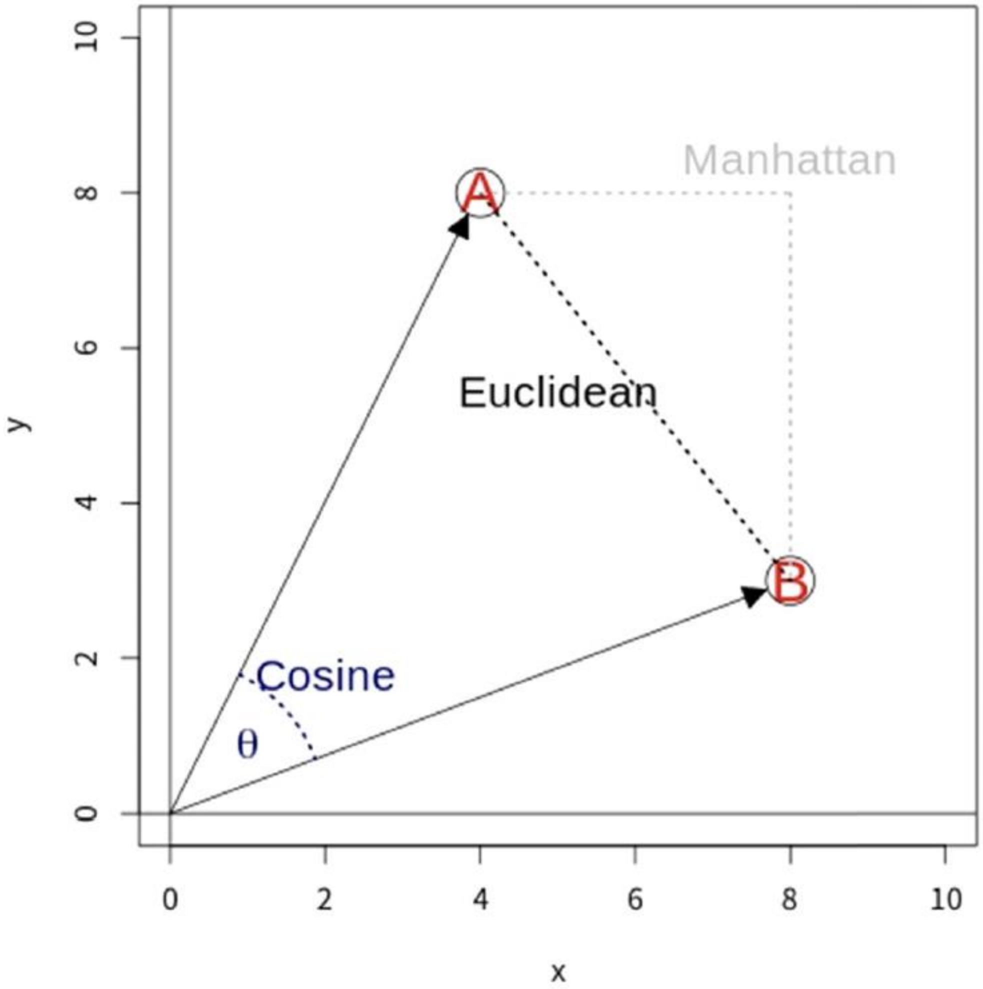
where *f* refers to a probability densitiy function, *x* and *y* the feature vectors of relative measured values (w_i^j, $i \in \{1 \ldots N\}, j \in (x, y\}$) of the documents of comparison (x = $\{w_1^x,\ \ w_2^x \ldots w_n^x\}$, and y = $\{w_1^y,\ \ w_2^y \ldots w_n^y)$).

For H_p and H_d, the probability density functions are required to be trained from the scores of the same-author and different-author comparisons respectively from the background database. Therefore, the proposition H_p is that the two documents of comparison (x, y) are written by the same author, while the proposition H_d is that the two documents of comparison (x,y) are written by different authors.

2.5.2.1 Cosine Distance Measure. For this thesis, the Cosine distance measure was used to calculate the scores of similarity. Figure 4 visualises the Cosine distance measure, along with the Euclidean and Manhattan distance measures, in a bi-dimensional space. The

Figure 4

Cosine Distance Measure with Euclidean and Manhattan Distance Measures in a Two-dimension Space



Note. Adapted from 'Score-based likelihood ratios for linguistic text evidence with a bag-of-words model' by S. Ishihara, 2021, *Forensic Science International, 327, 8.*

Cosine distance (D_c) concerns about only the angle θ of the feature vectors A and B, as in Figure 4, and not the scalar properties of the feature vectors. The Cosine distance can be mathematically expressed in Equation (9) as follows:

$$D_c(\mathrm{A, B}) = 1 - \cos(\theta) = 1 - \frac{\sum_i^n A_i B_i}{\sqrt{\sum_i^n A_i^2}\sqrt{\sum_i^n B_i^2}} \tag{9}$$

In Ishihara (2021), the three distance-based measures, namely the Euclidean, Manhattan, and Cosine distance measures, were used to calculate the scores of similarity of the documents of comparison which were generated from the same database. The fluctuation of the performance of the FTC system was being investigated as a function of the three different distance measures. Ishihara found that the Cosine similarly consistently worked best for the documents of any word length. Therefore, this thesis would use the Cosine similarty measure so as to attempt to obtain the most optimal results possible from the FTC system.

For this thesis, the function ‘t.cosine()’ of the ‘stylo’ R library was used to measure the Cosine similarity between two documents of comparison.

2.5.1.2 Probabilistic Distribution Models The probability density functions under H_p and H_d are required to be trained by the scores of the same-author and different-author comparisons from the background development, and such functions are used to convert scores into LRs. In this thesis, the distributions of scores were modelled using Normal, Log-normal, Gamma and Weibull distribution models, but the best-fitted model was selected via Akaike Information Criteria separately for same-author and different-author scores. The main reason why there were four distribution models used in this thesis is that the distributions of scores did not always conform to normality (i.e. Normal distribution; Ishihara, 2021). The other three distribution models, namely Log-normal, Gamma, and Weibull distributions, can better handle the skewed distributions of scores than Normal distribution.

2.5.3 Logistic-regression Calibration and Fusion

The LRC is a calibration of a set of LRs that may or may not be well calibrated by applying the linear shifting and scaling to those LRs based on the weights of the LRs derived from the development database. That is, the LRs of the same-author and different-author comparisons from the development database are used to train the LRC model, or specifically, the logistic regression line. The logistic regression line then applies linear shifting and scaling to the LRs of the testing database and derive the newly calibrated LRs. The aim of the LRC is straightforward: to minimise the magnitude of counterfactual LRs that support the incorrect hypotheses and contrast the correct hypotheses, and also to maximise the magnitude consistent-with-fact LRs that support the correct hypotheses.

The LRC can only calibration a single set of LRs; however, the logistic regression fusion (LRF) offers a calibration of multiple related sets of LRs (e.g. the LRs derived with SFs, WNGs, CNGs, and PNGs). The aim of the LRF is the same as outlined for the LRC: to minimise contrary-to-fact LRs and maximise consistent-with-fact LRs, which leads to a better FTC system performance. For the LRF, the logistic regression weight is calculated for each set of LRs; this is mathematically expressed in Equation (10):

$$log(fused\ likelihood\ ratio) = a_1x_1 + a_1x_1 + a_1x_1 + \ldots + a_nx_n + b \qquad (10)$$

where x_1, x_2, x_3, … x_n are the LRs of the first to the set of n, a_1, a_2, a_3, … a_n are the logistic regression weight for scaling, and b is the logistic regression weight for shifting. The logistic regression weight for both scaling and shifting is obtained from the development database. This thesis would be fusing the different numbers of WNGs, CNGs, and PNGs separately for each type of feature. After that, the fused FTC systems of every feature type (i.e. of WNGs, CNGs, and PNGs) and the best-performing combinations of SFs would eventually be fused together. If the FTC system perform gets better following the LRF, this means that the authorial information extracted from the different features or feature types are actually complementary. To apply the LRF, the FoCal Toolkit was used in R (Brümmer & du Preez, 2006). This thesis aims to empirically test whether the LRC is needed to further optimise the LRs derived via the score-to-LR conversion model (as specified in one of the research questions in Section 1.4).

2.5.4 Metrics of Performance

To assess the FTC system performance, log-likelihood-ratio cost (*Cllr*) is used in this thesis (Brümmer & du Preez, 2006). *Cllr* is calculated as in the following Equation (11):

$$Cllr = \frac{1}{2}\left(\left[\frac{1}{N_{SA}}\sum_{i}^{N_{SA}} log_2\left(1 + \frac{1}{LR_{SA_i}}\right)\right] + \left[\frac{1}{N_{DA}}\sum_{i}^{N_{DA}} log_2\left(1 + LR_{DA_i}\right)\right]\right) \qquad (11)$$

where the equation in the first square brackets evaluates the same-authors LRs, and that in the second square brackets evaluates the different-author LRs. In the first square brackets, N_{SA} refers to the number of the same-author comparisons, while LR_{SA_i} the LRs derived from the same-author comparisons. In the second square brackets, N_{DA} refers to the number of the different-author comparisons, while LR_{DA_i} refers to the LRs derived from the different-author comparisons.

Cllr penalizes all the likelihood ratios except zero and + infinity. Furthermore, the contrary-to-fact LRs are penalised more than the consistent-with-fact LRs, and the extent of the penalty will increase as the contrary-to-fact LRs move father away from unity (i.e. one). For example, the contrary-to-fact same-author LR of -10 would be penalised more than the consistent-with-fact same-author LR of 10. However, the contrary-to-fact different author LR of 30 would be penalised more than the contrary-to-fact different-author LR of 20 since the former is more misleading than the latter. Note that for a perfect FTC system, the consist-with-fact same-author LRs would exceed unity (>1) while the consistent-with-fact different-author LRs would go below unity (<1). The penalties of all the likelihood ratios are accumulated into one single *Cllr* value for each feature set or experiment setting; the lower the *Cllr* value is, the better the FTC system performance will be.

Cllr further comprises two sub-components: a discrimination loss (*Cllr*^{min}) and a calibration loss (*Cllr*^{cal}). *Cllr*^{min} is the minimum *Cllr* value gained after applying a pooled-adjacent-violators (PAV) transformation to the derived LRs, while *Cllr*^{cal} is calculated by subtracting the *Cllr*^{min} from the *Cllr*. In this thesis, the *Cllr*^{min} value is used to investigate the discrimination power of each feature set or experiment setting; the lower the *Cllr*^{min} value, the less the loss of the discrimination power is and the better the discrimination power will be. The *Cllr*^{cal} is used to investigate the calibration performance of each feature set or experiment setting; the lower the *Cllr*^{cal} value is, the less the loss in calibration performance is and the calibration performance will be. In calculating these metrics of performance for this thesis, the FoCal Toolkit in R was used (Brümmer & du Preez, 2006).

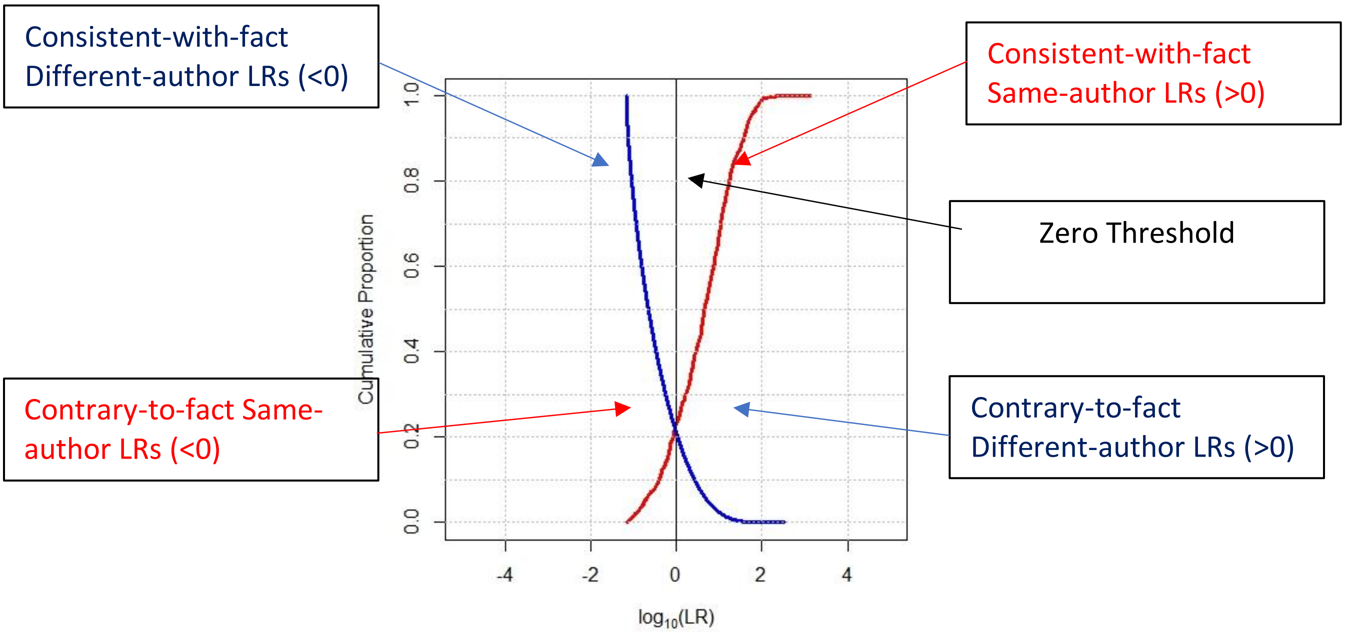
2.5.5 Tippett Plots

To visually present and assess the quality of the LRs derived, Tippett plots are employed in this thesis. Figure 5 displays an example of a Tippett plot which can tell us a lot of things regarding the quality of the LRs derived.

Considering Figure 5, what is the most obvious is the magnitude of the same-author LRs and the different-author LRs. As mentioned earlier, the FTC system will work perfectly if the same-author LRs exceed unity and the different-author LRs go below unity. However, since the LRs are presented in the log10 scale, unity has shifted from one to zero (log10(1) = 0). The magnitude of the consistent-with-fact same-author LRs (represented in red line right to the zero threshold) is approximately the magnitude of 3, while the magnitude of the consistent-with-fact different-author LRs (represented in blue line right to the zero threshold) is approximately the magnitude of -1 only. This means that this feature set or experiment setting provides a stronger support for the same-author hypothesis than for the different-author hypothesis. However, there is also the noticeable magnitude of contrary-to-fact LRs for both same-author (approximately -1.5) and different-author comparisons (approximately 2.5). All the LRs presented all contribute to the *Cllr* value of the experiment setting of this Tippett plot, with the contrary-to-fact LRs being more penalised and therefore resulting in the increase of the *Cllr* value more than the consistent-with-fact LRs.

Figure 5

Example of Tippett Plot Visually Presenting Quality of Derived LRs



Note. For Tippett plots presented in this thesis, the red line represents the LRs derived from the same-author comparisons, while the blue line represents the LRs derived from the different-author comparisons. For the LRs derived from the same-author comparisons, the LRs beyond the zero threshold (>0) are the consistent-with-fact LRs, while the LRs below the zero threshold are the contrary-to-fact LRs (<0). For the LRs derived from different-author comparisons, the LRs below the zero threshold are the consistent-with-fact different-author LRs, while the LRs beyond the zero threshold are the contrary-to-fact different-author LRs.

We can also see whether the LRs derived are well calibrated or not by considering Tippett plots. From Figure 5, the crossing point between the red and blue lines is visually aligned with the zero threshold; this means that the LRs derived presented in this Tippett plot are well-calibrated. Moreover, the crossing point reflects the $Cllr^{cal}$ value, which tells us about the calibration performance: the lower the $Cllr^{cal}$ value is, the better the calibration performance will be.

Chapter 3 Logistic Regression Calibration: Results and Discussion

The aim of this chapter is to investigate whether the LR-based FTC system further needs another calibration, as known as logistic regression calibration. I will also attempt to discuss why the logistic regression is or is not necessary in the LR-based FTC system.

To see whether the LRC is necessary for the FTC system and whether the LRs needs to be calibrated through a second calibration, I will display the results of the best-performing pairs of statistical features (SFs) before and after the LRC in Table 2. What stands out from Table 2 is that the pre-LRC and post-LRC C_{llr} values of the same pairs of features are very similar to each other for every document length. As mentioned in Section 2.5.4, the closer to zero the C_{llr} value gets, the better the FTC system performance will be. This means that the LRC brings little improvement, or even deterioration in some cases, in the C_{llr} values and thus the FTC system performance.

In assessing how well LRs are calibrated in LR-based FTC, as important as the C_{llr} values are the C_{llr}^{cal} values, which are used to quantify calibration loss. The closer to zero the C_{llr}^{cal} value gets, the better the calibration performance will be. As observed from the Table 2, much like the case of the *Cllr* values, the pre-LRC and post-LRC C_{llr}^{cal} are much alike, suggesting that the LRC does not bring any significant change to the calibration performance. Furthermore, considering the pre-LRC the C_{llr}^{cal} values, they are already very close to the zero threshold. This suggests that by using only the first calibration which is via the score-to-LR conversion model, the FTC system has already been well calibrated.

Another way to manifest that the second calibration, or the LRC, is unnecessary is by visual comparison of pre-LRC and post-LRC LRs. This can be done by using Tippett plots (see Section 2.5.5). Figure 6 displays the Tippett plots of the LRs derived with the same-author and different-author comparisons using only the best-performing pair for each

Table 2

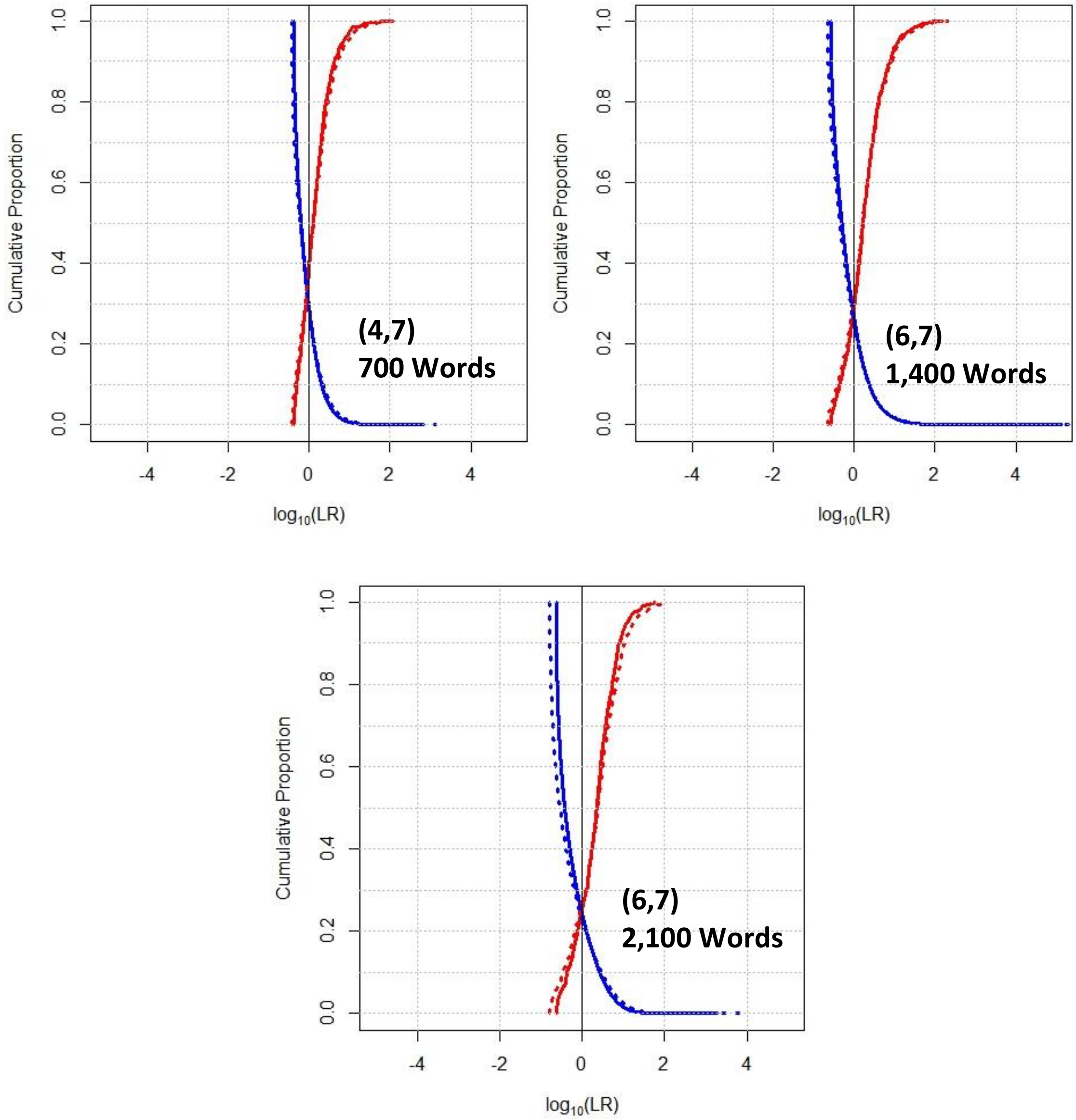
Pre-Logistic Regression Calibration and Post-Logistic Regression Calibration Results of the Best-Performing Pairs of Statistical Features Separately for Documents of 700, 1,400, and 2,100 words

Sample Word Number	Pair of Features	C_{llr}		C_{llr}^{cal}	
		Pre-LRC	Post-LRC	Pre-LRC	Post-LRC
700	(4,6)	0.909	0.907	0.022	0.021
	(6,9)	0.917	0.918	0.024	0.025
	(4,7)	0.934	0.931	0.024	0.021
	(4,9)	0.937	0.939	0.027	0.028
	(7,9)	0.938	0.937	0.032	0.019
	(1,6)	0.945	0.940	0.029	0.024
	(3,6)	0.945	0.942	0.024	0.020
	(1,4)	0.948	0.945	0.022	0.019
	(1,7)	0.953	0.950	0.025	0.022
	(1,9)	0.955	0.954	0.019	0.017
1,400	(6,7)	0.839	0.838	0.028	0.027
	(4,6)	0.840	0.833	0.033	0.025
	(6,9)	0.858	0.858	0.035	0.034
	(1,6)	0.867	0.867	0.033	0.033
	(4,7)	0.876	0.868	0.027	0.019
	(7,9)	0.877	0.875	0.030	0.028
	(1,7)	0.887	0.888	0.029	0.030
	(3,6)	0.888	0.887	0.029	0.028
	(1,4)	0.891	0.891	0.022	0.022
	(4,9)	0.893	0.893	0.035	0.035
2,100	(6,7)	0.777	0.776	0.036	0.036
	(4,6)	0.792	0.787	0.027	0.022
	(6,9)	0.801	0.773	0.801	0.773
	(1,6)	0.809	0.809	0.026	0.026
	(4,7)	0.826	0.821	0.024	0.019
	(7,9)	0.829	0.826	0.033	0.030
	(3,6)	0.831	0.829	0.028	0.026
	(1,7)	0.840	0.839	0.022	0.021
	(3,7)	0.852	0.852	0.028	0.027
	(1,4)	0.856	0.856	0.021	0.022

Note. Pre-LRC stands for ‘pre-logistic regression calibration’; Post-LRC stands for ‘post-logistic regression calibration’. The combinations rank based on their *Cllr* values; the higher in rank they are, the lower the *Cllr* value and the better the FTC system performance is. For more information on how to interpret *Cllr* and *Cllr*^{cal}, see Section 2.5.4

Figure 6

Tippett Plots of Best-performing Pairs of Statistical Features Separately for Documents of 700, 1,400, and 2,100 words



Note. Solid lines represent pre-LRC LRs, dotted lines represent post-LRC LRs, red lines represent same-author LRs; blue lines represent different-author LRs. For more information on how to read Tippett Plots, see Section 2.5.5

document length. It can be observed in Figure 6 that the dotted lines, representing the pre-LRC LRs, and the solid lines, representing the post-LRC LRs, differ only to a small extent as they are almost aligned with each other. Not only that, those intersections of the pre-LRC LRs are also neatly aligned with the zero threshold. This coincides with the C_{llr}^{cal} values of the pre-LRC FTC system that are already very close to zero. These facts indicate that pre-LRC FTC system has already been well calibrated and does not need another calibration.

These facts help emphasise that the pre-LRC LRs, are only minimally impacted by the LRC calibration even if the impact is favorable and thus make the LRC unnecessary. In conclusion, there has been proof in several ways that the LRC rarely brings any significant improvement, or in some cases, deterioration, to the FTC system performance. This suggests that the LRs derived from the same-author and different-author comparisons of the pre-LRC are ready to be interpreted as the weight of evidence.

So far, I have only presented the results of the LRC on the FTC system with two SFs. That said, according to my observation of the results of the LRC on the other SFs and also the other feature types, the minimal impacts of the LRC as described in this chapter well resonate across all the features. That is, the LRC is deemed to be unnecessary for all the FTC experiments conducted for this thesis, irrespective of the features or features types and of the dimension of the feature vectors on which the LRC is tested

Such impacts of the LRC on the FTC system does not strike as a surprise. Although Carne and Ishihara (2020) and Ishihara (2014, 2017a, 2017b) demonstrated that the LRC was actually needed to calibrate the scores obtained, this need of the LRC occurred because of the approaches they used (i.e. the feature-based and similarity-typicality score-based approaches). In this thesis, the similarity-only score-based method, unlike the other two said approaches, converts the multidimensional feature space to a univariate score space. Since it is univariate, it can be appropriately modelled even with a limited amount of data. In the similarity-only-score-based method, if there is a decent amount of data (e.g. the database used in this thesis) that can be used to train the score-to-LR conversion model, the score-to-LR conversion model should return well-calibrated LRs. Previous studies also demonstrated that the similarity-only score-based method yielded well-calibrated LRs (Block et al., 2015; Garton et al., 2020). Based on LR-based FTC, Ishihara (2021) also showed that the similarity-only score-based approach yielded well-calibrated LRs and the LRC was not needed.

Since it is already known that the FTC system built for the purposes of this thesis does not need to be calibrated again via LRC, from this point onwards, I would only consider and present the pre-LRC LRs which will be assessed via the C_{llr} values in the following experiments conducted for each type of features.

Chapter 4 Statistical Features: Results and Discussion

The aim of this chapter is to present the results of the experiment on using statistical features (SFs) to capture writing idiosyncrasies via likelihood ratios (LRs) and then explore meaning of those numerical results in relation to how those results can imply subtleties of individuals' writing styles. In this chapter, I will be presenting the best-performing feature combinations for each type of stylometric features in tables and visualising how the FTC system performance changes as a function of feature numbers. Then I will be disccusing the results' implications to authors' differing writing styles.

In this thesis, I will only present the most important tables of results in the main text, so findings presented here will be as concise as possible. I will be putting full results of all the experiments in the Appendices of this thesis for the reader's reference. For full results of SFs, see Appendix A.

4.1 Best-performing Statistical Features

Table 3 summarises the best-performing combinations of SFs across all the combinations for each document length with their associated C_{llr} and C_{llr}^{min}, values. The combinations in Table 3 rank according to the C_{llr} values for each document length. All the combinations displayed in Table 3 has one thing in common: they all feature the SFs number 4 (FK), number 6 (PuncRatio), number 7 (SpecialCharRatio), number 8 (UpperCaseRatio), and number 9 (CharNumberPerWord). This makes it safe to conclude that these five features are the best performing SFs across all the combination points and all the document lengths. Three other SFs that perform well when added to the existing five best performing features, especially in the document lengths of 1,400 and 2,100, are the SFs number 1 (TTR), number 2 (K), and number 5 (DigitRatio). Meanwhile, the SF number 10 (FreqUnusualWordUs) seems to be the worst-performing SF among the ten SFs tested in this FTC system as it appears only two or three times in the bottom list of the document lengths of 1,400 and 2,100,

Table 3

Best-performing Combinations of Statistical Features Separately for Documents of 700, 1,400, and 2,100 words

Sample Word Number	Combination of features	*Cllr*	*Cllr*^{min}
700	(3,4,6,7,8,9)	0.814	0.792
	(4,6,7,8,9)	0.814	0.791
	(1,4,6,7,8,9)	0.816	0.791
	(1,4,5,6,7,8,9)	0.817	0.797
	(3,4,5,6,7,8,9)	0.818	0.799
	(2,3,4,5,6,7,8,9)	0.819	0.801
	(1,2,4,5,6,7,8,9)	0.819	0.8
	(1,2,4,6,7,8,9)	0.82	0.8
	(2,3,4,6,7,8,9)	0.821	0.804
	(4,5,6,7,8,9)	0.822	0.8
1,400	(1,2,4,5,6,7,8,9)	0.669	0.652
	(1,2,4,6,7,8,9)	0.67	0.648
	(1,4,5,6,7,8,9)	0.671	0.655
	(1,4,6,7,8,9)	0.671	0.65
	(2,3,4,5,6,7,8,9)	0.673	0.653
	(2,3,4,6,7,8,9)	0.674	0.656
	(1,2,4,5,6,7,8,9,10)	0.678	0.664
	(3,4,5,6,7,8,9)	0.68	0.662
	(2,3,4,5,6,7,8,9,10)	0.681	0.665
	(1,2,4,6,7,8,9,10)	0.681	0.663
2,100	(1,4,6,7,8,9)	0.594	0.571
	(1,4,5,6,7,8,9)	0.595	0.575
	(1,2,4,6,7,8,9)	0.596	0.575
	(1,2,4,5,6,7,8,9)	0.597	0.575
	(4,6,7,8,9)	0.601	0.578
	(2,3,4,6,7,8,9)	0.601	0.579
	(1,2,4,5,6,7,8,9,10)	0.601	0.581
	(2,3,4,5,6,7,8,9)	0.601	0.579
	(1,2,4,6,7,8,9,10)	0.602	0.584
	(3,4,5,6,7,8,9)	0.603	0.582

Note. The combinations rank based on their *Cllr* values; the higher in rank they are, the lower the *Cllr* value and the better the FTC system performance is. For more information on how to interpret *Cllr* and *Cllr*^{min}, see Section 2.5.4

Table 4

Results of All the Ten Statistical Features Combined as a Benchmark Comparison to Best-performing Combinations of Statistical Features Separately for Documents of 700, 1,400, and 2,100 words

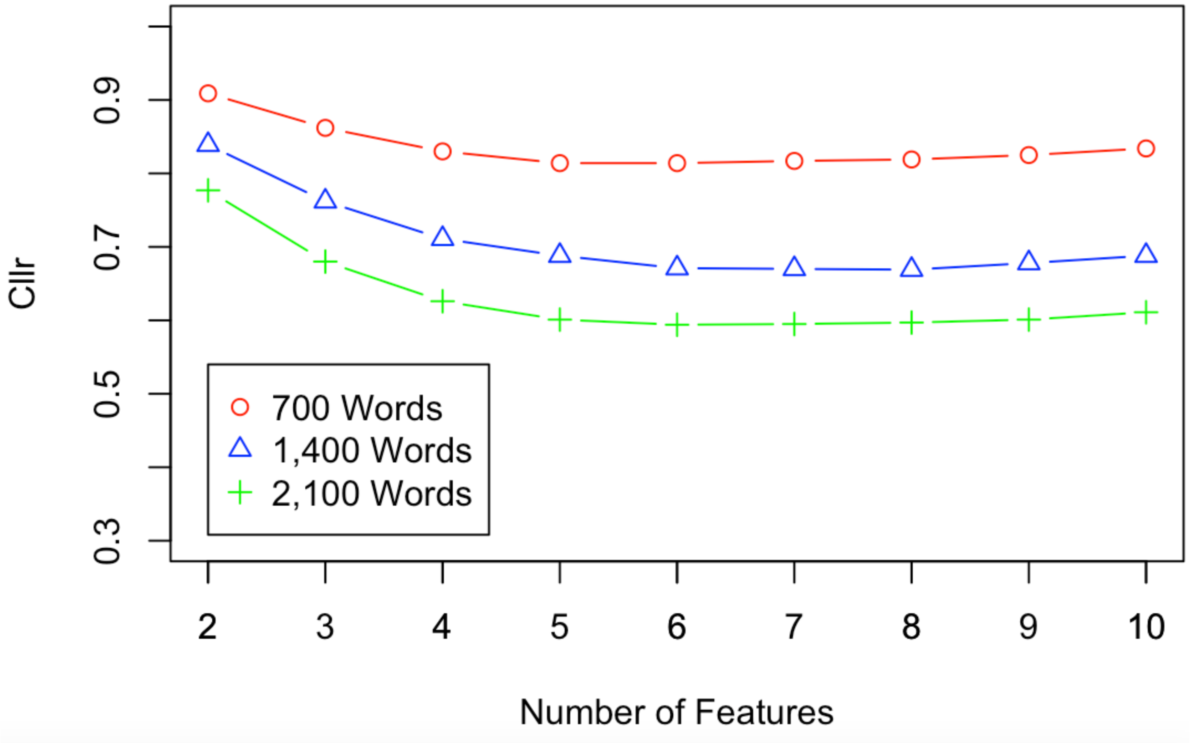
Document length	C_{llr}	C_{llr}^{min}
700	0.834	0.819
1,400	0.688	0.671
2,100	0.611	0.593

and does not appear at all in the document length of 700. It is worth noting that the SF number 3 (Hapax), although yielding mediocre results when used under the document length of 1,400 and 2,100, works considerably well when used under the document length of 700, as reflected in it being part of the combination (3,4,6,7,8,9), the best performing SF combination of the document length of 700.

From Table 3, there is no trace of all the ten SFs combined as the best-performing combinations for SFs. This means that it does not take all the available SFs for the FTC system to get optimal results. Table 4 displays the C_{llr} and C_{llr}^{min}, values of the combinations of all the ten SFs tested at the same experiment setting. It can be observed that the C_{llr} values of the best-performing SF combination in Table 3 are less than those of all the ten SFs in Table 4, with the *Cllr* discrepancies of 0.020, 0.019, and 0.017, for the documents of 700, 1,400, and 2,100 words respectively. The discriminating power, as reflected in C_{llr}^{min}, of those ten-feature combinations in Table 4 are also worse than those best-performing combinations in Table 3, with the $Cllr^{min}$ discrepancies of 0.027, 0.019, and 0.022 for the documents of 700, 1,400 and 2,100 words respectively. In fact, as in reflected in Table 3 and Table 4, one can even surpass the level of results of all the ten SFs as soon as there are only

Figure 7

Visualisation of the Change of C_{llr} Values as a function of the Number of Statistical Features Separately or Documents of 700, 1,400, and 2,100 words



five features in the experiment. For example, for the documents of 700 words, while the best-performing combination of all the ten SFs has the C_{llr} value of 0.834, the five-feature combination (4,6,7,8,9)—the second-to-best-performing combination—performs even better than the ten-feature combination, boasting the C_{llr} value of 0.814.

4.2 Performance of Statistical Features as a Function of the Number of Features

Figure 7 presents how the increase or decrease in the number of SFs correlates with the C_{llr} values of the best-performing combinations for each combination number of SFs. For the documents of 700 words, the FTC system performance starts to hardly get any better when there are six SFs at the same experiment setting as the C_{llr} value does not change (C_{llr} = 0.814) when the number of SFs in one feature combination increases from five to six. The FTC system performance starts to deteriorate in a gradual fashion when there are seven or more SFs in the FTC system.

For 1,400 words, the shift from eight to nine SFs is where the FTC system performance gets saturated as the C_{llr} value increases from 0.669 to 0.678, and the FTC system performance starts gradually deteriorating afterwards.

For 2,100 words, it is not until there are eight SFs in that the FTC system performance gets saturated since there is the increase in the C_{llr} value from 0.595 to 0.597 when the number of SFs increases from seven to eight. As with the other two document lengths, the FTC system performance continually deteriorates after performance saturation for 2,100 words.

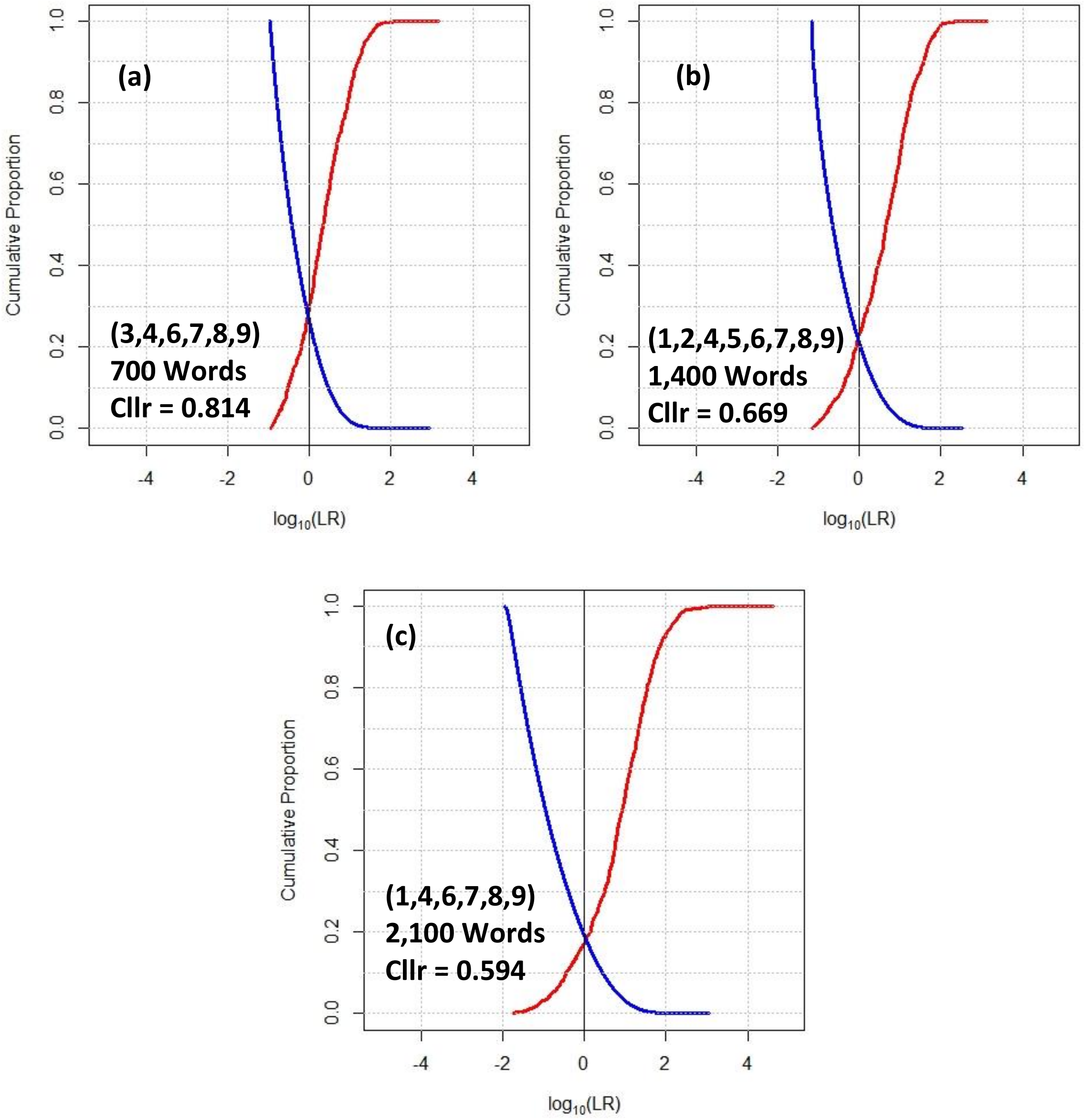
4.3 Tippett Plots of the LRs Derived with Statistical Features

Figure 8 illustrates the Tippett plots of the best-performing combinations of the SFs for each document length. The six-feature combination (3,4,6,7,8,9) works best for the document length of 700, boasting the C_{llr} value of 0.814. For 1,400 and 2,100, the eight-feature combination (1,2,4,5,6,7,8,9) (C_{llr} = 0.669) and the six-feature combination (1,4,6,7,8,9) (0.594) work best respectively. As reflected in the magnitude of the consistent-with-fact same-author LRs surpassing that of the consistent-with-fact different-author LRs, the experiments using a series of SFs provide stronger support for the same-author hypothesis more than they do for the different-author hypothesis. For example, considering Figure 4(c), the strongest consistent-with-fact same-author LR is 4.619, but the strongest consistent-with-fact same-author LR is only -1.960.

However, considering Figure 8, this FTC system using only SFS is still not perfect as there is the considerable magnitude of the contrary-to-fact same-author LRs (represented in red lines left to the zero threshold) and the contrary-to-fact different-author LRs (represented in blue lines right to the zero threshold) in every combination. The magnitude of the contrary-

Figure 8

Tippett Plots of Best-performing Combinations of Statistical Features for Each Document length (700, 1,400, and 2,100 Words)



to-fact different-author LRs is even greater than that of the consistent-with-fact different author LRs. These contrary-to-fact LRs serve to prove that there are still some errors which contradict the correct hypotheses, especially in the different-author comparisons, and it therefore leaves room for improvement. Approaches to the improvement of LR-based FTC in terms of counterfactual LRs may be a subject of future study.

4.4 Discussion

Now it is known from the SF experiment (see Section 4.1) that if one selects some certain features—for example the Flesch-Kincaid readability scores, the ratio of punctuation marks, the ratio of special characters, etc—the FTC performance tends to yield good results than using the other SFs. I will be discussing these best-performing features in the follow discussion.

The first two well-performing SFs that will be discussed are the Flesch-Kincaid readability scores and the average number of characters per word. As explained in Section 2.4.1, the Flesh-Kincaid readability scores determines how hard it is to read the text. It considers the average length of sentence and the number of syllables and words; the longer the text is, the harder the text can be read. This makes the Flesch-Kincaid readability scores and the average number of characters per word quite alike in the sense that the longer an author writes, the more likely the FTC system will be able to discriminate the author's writing from the others' writings. However, these two features are different in their equations, and one has already been tested in LR-based FTC while another has not. My speculation regarding the usefulness of the Flesch-Kincaid readability scores and the average number of characters per word is that some authors tend to write more difficult, longer sentences than others regardless of topics or platforms on which they are writing. There are certainly some authors who like to keep it short and simple irrespective of writing settings. Therefore, my speculation is that it would not be such a daunting task to discriminate either a difficult text which contains a lot of long sentences and long words or a surprisingly short text from texts of normal difficulty and length.

The other three best-performing SFs are the ratios of punctuations, special characters, and uppercase characters. The calculation of these SFs is straightforward: the accumulation of these features is divided by the length of sentence within a specific feature type. Thus, according to the calculation of these ratios, if an author uses these features more or less than usual in their writing, the FTC will be able to discriminate that writing from the others more easily than if the author uses these features in a conventional manner as with the other authors. My argument for the usefulness of these features is that some writers have a fixated way of using punctuations, special characters, and uppercase characters in their writings. For example, some authors, especially when they write in an informal manner, write in lowercase in most of their writings. Some have a habit of frequently or infrequently using punctuation marks and special characters, and sometimes this may be a result of another writing habit. For instance, those who write in short sentences tend to end up with more periods than those who write in long sentences. Another example is that when there is a need to exemplify what has been written prior, some authors tend to use parentheses rather than commas.

There are also some SFs that perform not as well as the five well-performing SFs that have been discussed earlier, especially hapax legomenon (i.e. the frequency of words that appear only once) for the documents of 1,400 and 2,100 words, and also the frequency of unusual words. Despite featuring in the best-performing SF combination (3,4,6,7,8,9) for the documents of 700 words, hapax legomenon, the SF that has yet to be tested in LR-based FTC, produces subpar performance compared to the other SFs. The possible explanation for this phenomenon is that for the documents of longer length (i.e. 1,400 and 2,100 words), hapax legomena have more chance to appear more than once than for the documents of shorter length (i.e. 700 words). When there are less hapax legomena, the ratio of hapax legomena are supposed to be low across the documents of lower length and regardless of topics, making hapax legomenon not as useful as it is for the documents of shorter length. On the other hand, when there are more hapax legomena, there is more chance for its ratio to be more varied across the documents of shorter length, making hapax legomena a useful SF for the documents of 700 words.

For the frequency of unusual words, this comes as a surprise since it has been proved to work considerably well in FTC within the LR paradigm despite being calculated by different spelling dictionaries and on different writing genres and platforms (Ishihara, 2017a, 2017b). My speculation as to the unusefulness of the frequency of unusual words is that as the documents tested in the FTC system are product reviews, they tend to be written in a casual manner and not to consist of many unusual words compared to text of more formal genres. That said, Ishihara (2017a, 2017b) experimented on LR-based FTC on chatlog messages, which can be argued to fall in the informal end of writing genres. One possible explanation is that in chatlog messages, it is more likely than in product reviews for authors to write unusual words by the standard of a spelling dictionary, which can include personal information, abbreviations, acronyms, slangs, etc. In writing chatlog messages, when an author overtly or scarcely used unusual words, it may be easier to single out that author from the rest than in writing product reviews.

Within LR-based FTC, Ishihara (2017a, 2017b) made use of lexical features to conduct distance-based FTC experiments in chatlog messages and found that the most outstanding features are the type-token ratio (TTR) and Yule’s K, while one of the other well performing features is the ratio of digits. This is somewhat contrasting to my findings: TTR and Yule’s K are not ones of the most robust SFs in the FTC system with which I was working, and they are best used as an addition to the existing combinations of the well-preforming SFs. Regarding the ratio of digits, it functions well when it is tested along with the aforementioned robust SFs, but it is not one of the most robust SFs in my FTC system. I reckon that this again has to do with different writing platforms on which the FTC systems of this thesis and Ishihara (2017a, 2017b) were built. Chatlog messages are likely to be shorter than product reviews, which can determine the discriminating power of vocabulary richness measures (i.e. TTR, Yule’s K, Honoré’s R) differently from product reviews. Chatlog messages are also expected to include more digits (e.g. personal information, time, etc.), leading to the ratio of digits being more powerful for chatlog messages than for product reviews. These results from my thesis can also suggest that some certain features (e.g. the frequency of unusual words, the vocabulary richness measures, etc.) are more suitable to use in an FTC task on one-to-one communication than on one-to-many communication as is the case for this thesis.

The fact that the longer document takes more SFs in order to reach its potential also warrants some discussion. To my understanding, the documents of shorter length (i.e. 700 words) do not provide a great amount of data for the FTC system to work with. This means that short documents are likely to be saturated in the FTC system performance more quickly than long documents. On the other hand, the documents of longer length (i.e. 1,400 and 2,100 words) are more auspicious than those of shorter length in terms of an amount of data. Consequently, long documents are open for possibilities for the reoccurrences of features more than short documents, which leads to that it takes longer for long documents to reach a performance saturation point.

Chapter 5 Word N-gram Features: Results and Discussion

The aim of this chapter is to present the results of the FTC system built by using only word n-gram features (WNGs) to derive same-author and different-author likelihood ratios (LRs) for different experiment settings. In this chapter, I will be presenting the best-performing experiment settings using WNGs in tables and visualising how the FTC system performance changes as a function of feature numbers. After that, I will be presenting the results of the fused FTC system where all the numbers of WNGs are fused together. Lastly, I will also discuss the significance of the numerical results where necessary. For full results of the experiment settings using WNGs at each number of WNGs, see Appendix B.

5.1 Best-performing Word N-gram Features

Table 5 displays the best-performing experiment settings of WNGs. The most notable is that word unigrams (n=1, WN1s) overwhelmingly excel for every document length with only several appearances of word bigram (n=2; WN2s) in the bottom half for 1,400 and 2,100 words. When they reach their potential, the best-performing experiment settings of WNGs, which are all WN1s, yield the *Cllr* values of 0.711, 0.438, and 0.302, and the $Cllr^{min}$ values of 0.695, 0.424, and 0.292 for the documents of 700, 1,400, and 2,100 words respectively. What is also interesting here is that the numbers of word n-gram features that perform best are not around the proximity of 600 features—which are the maximum number of features tested for the WNG experiments—but instead are the numbers in the middle range. As can be observed in Table 3, 300 WNGs yield the most optimal result for every document length, and the middle numbers, such as the likes of 275, 325, 350, and 375 WNGs, dominate the top half of the best-performing WNGs for each document length. This is once again the proof that sometimes we do not need all available features tested at the same experiment setting to have the most optimal results possible.

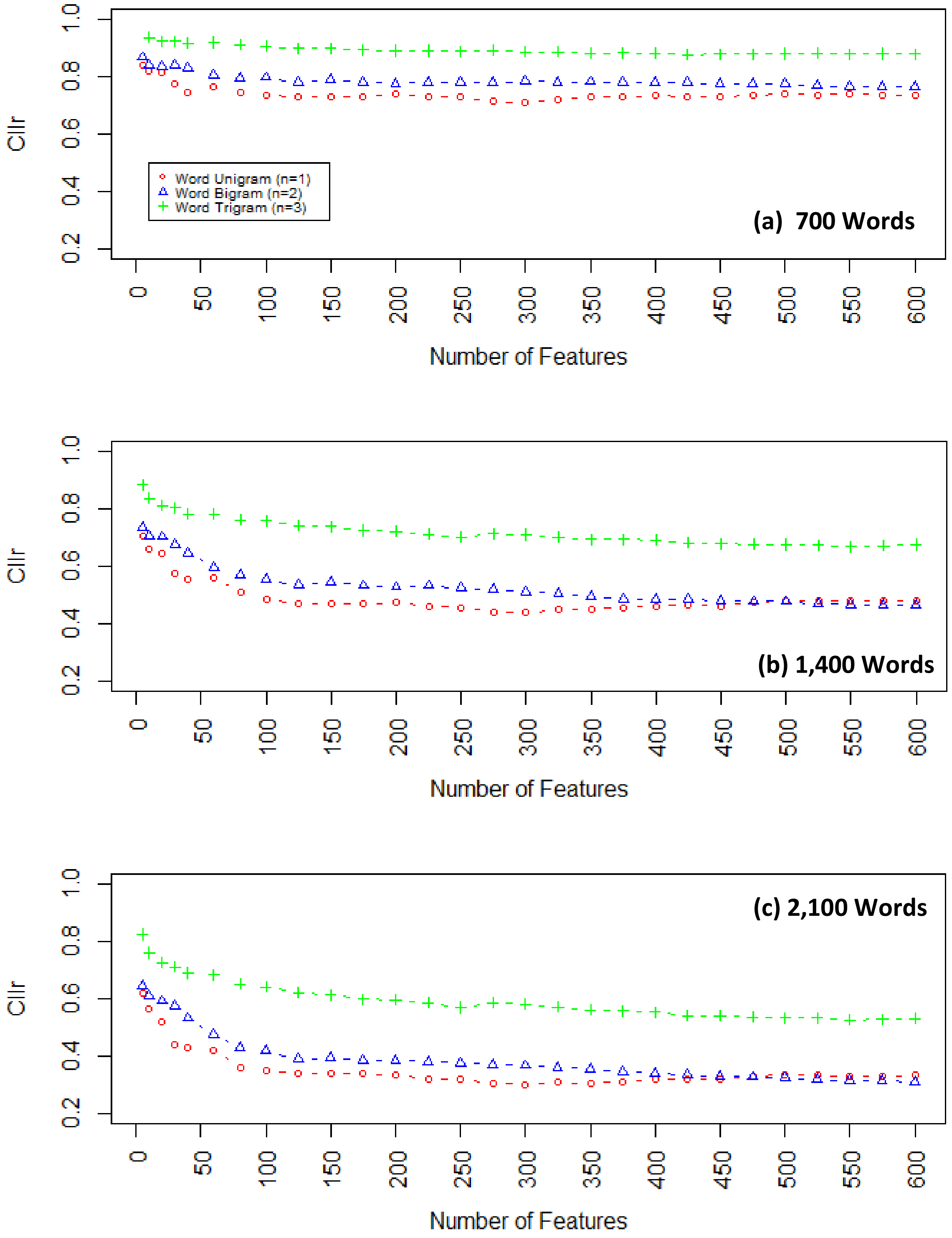
Table 5

Best-performing Experiment Settings of Word N-gram (n=1,2,3) Features for 700, 1,400, and 2,100 Word Lengths

Sample Word Number	Number of N-gram	Number of Features	*Cllr*	*Cllr^{min}*
700	1	300	0.711	0.695
	1	275	0.712	0.697
	1	325	0.721	0.704
	1	350	0.728	0.711
	1	225	0.728	0.712
	1	250	0.728	0.716
	1	125	0.728	0.714
	1	450	0.729	0.712
	1	375	0.729	0.711
	1	175	0.729	0.715
1,400	1	300	0.438	0.424
	1	275	0.441	0.427
	1	325	0.449	0.435
	1	350	0.452	0.441
	1	375	0.457	0.446
	1	250	0.457	0.443
	1	400	0.46	0.445
	1	225	0.461	0.448
	1	450	0.462	0.45
	2	575	0.464	0.453
2,100	1	300	0.302	0.292
	1	275	0.305	0.294
	1	350	0.308	0.297
	1	325	0.308	0.298
	1	375	0.311	0.300
	2	600	0.312	0.300
	2	575	0.314	0.300
	2	550	0.314	0.301
	2	525	0.319	0.305
	1	400	0.318	0.307

Figure 9

C_{llr} Values Plotted against Number of Features of Word Unigrams (n=1), Word Bigrams (n=2), and Word Trigrams (n=3) for 700, 1,400, and 2,100 Word Lengths



5.2 Performance of Word N-gram Features as a Function of the Number of Features

Although WN1s perform best, it is wise to look at how well WN2s and word trigrams (WN3s) perform compared to word unigrams; this is visualised in Figure 9. First off, the FTC system performances using different numbers of WNGs fluctuates to a certain degree for 1,400 and 2,100 words, while the FTC system performs more consistently when the word length is 700 words. WN2s and WN3s also take more features than WN1s to reach their peaks. For WN1s, it is known from Table 5 and Figure 9 that the number of features that yields the most optimal results is 300 words across every document length. For WN2s and WN3s, the numbers of features that yield the best results fall near the maximum end of the number of features tested, which are 600 features, as seen in Figure 9(b) and Figure 9(c). Unlike WN1s, as the number of features increases, WN2s and WN3s rarely drop in performance as they have already taken longer than word unigrams to hit their performance ceilings, especially for the documents of 1,400 and 2,100 words. For the documents of 1,400 words, the FTC system performance of WN2s even starts to overtake that of WN1s when there are roughly 475 to 500 features in the FTC system. This also reinforces my stance that WN1s cannot effectively handle the higher numbers of features as well as the higher numbers of WNGs.

5.3 Fused FTC System between Word N-gram Features

The different nature of the different WNGs (n=1,2,3) tested for this thesis might extract different pieces of authorial information from text. Those bits of authorial information may even complement each other and build up stronger support for authorship hypotheses. Provide that this is true, the fusion of the LRs derived with different WNGs will be expected to improve the FTC system performance. To investigate whether this is true, the LRs derived with different word n-grams were logistic regression fused (LRF).

Table 6

Results of Fused FTC System Using Word N-gram Features (n=1,2,3; Top Table) in Comparison to Best Performing Experiment Settings of Word N-gram Features (Bottom Table) for 700, 1,400, and 2,100 Word Lengths

Sample Word Number	*Cllr*	$Cllr^{min}$
700	0.665	0.651
1,400	0.363	0.349
2,100	0.230	0.219

Sample Word Number	Number of N-gram	Number of Features	*Cllr*	$Cllr^{min}$
700	1	300	0.711	0.695
1,400	1	300	0.438	0.424
2,100	1	300	0.302	0.292

Table 6 presents the results of the fused FTC system using three different WNGs in comparison to those of the best-performing combinations of WNGs adapted from Table 5. Overall, the fused FTC system performs considerably better than its corresponding FTC systems, especially when the length of the documents in comparison is long. For the documents of 700 words, the C_{llr} value of the fused system is better than that of the best-performing experiment setting of WNGs by a C_{llr} of 0.046. Nonetheless, for the documents of longer length (1,400 and 2,100 words), the C_{llr} values of the fused system beat those of the corresponding best-performing settings by a C_{llr} of 0.075 and 0.072 respectively. The dricriminating power, as reflected in the $Cllr^{min}$ values, also see an improvement following the LRF, with the $Cllr^{min}$ improvements of 0.044, 0.075, and 0.085 for the documents of 700, 1,400, and 2,100 words respectively. The fact that the LRF leads to the better FTC system performance has proved that different numbers of WNGs can tell us distinctive yet complementary authorial information on written text.

The success following the LRF implies the strong complimentary relations between the three different numbers of WNGs. To explain, although WN1s manage to give us the best results for WNGs, they cannot capture sequential information in the way WN2s and WN3s do since WN1s are the orderless presentations of text (i.e. the presentations where only the most frequent WNGs and their relative frequencies are extracted). With the assistance of the LRF, the fused FTC system may take advantage of the high discriminating power of WN1s and the order-minded representation of text of WN2s and WN3s.

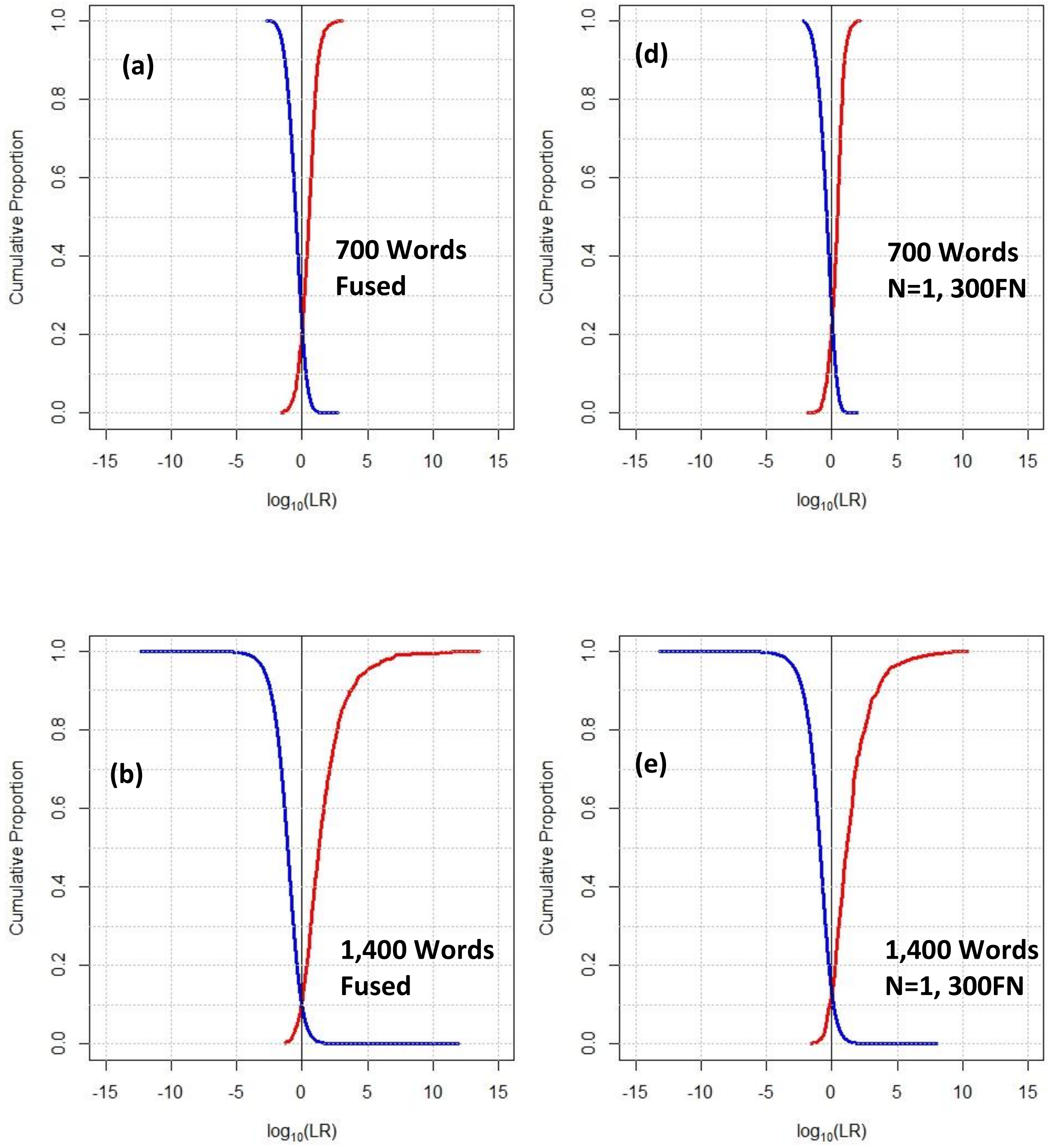
5.4 Tippett Plots of Fused FTC System between Word N-gram Features

Figure 10 displays the Tippett plots of the fused FTC system with the three different WNGs (n=1,2,3) in comparison to the FTC system with each of the n-gram which performs the best for each document length. Visual inspection may not say much about the changing magnitude of LRs before and after fusion, especially for the documents of 700 words. That said, a closer visual inspection reveals that when the FTC system performance improves because of the LRF and the C_{llr} value increases, the magnitude of the fused LRs is seen to be somewhat bigger than before the LRF. For example, for the documents of 2,100 words, the magnitude of the fused same-author LRs (Figure 10(f)) is seen to be farther away from the threshold of log10LR of 0 (i.e. exceeding the magnitude of log10LR of 15) than the same-author LRs of the best-performing experiment setting in the pre-fusion FTC system (Figure 9(c)); this is reflected in the *Cllr* discrepancy of 0.072 and the $Cllr^{min}$ discrepancy of 0.073.

However, there is still the same problem as in Section 4.3 occurs here as well. That is, the bigger magnitude of the consistent-with-fact LRs comes with the cost of the bigger magnitude of the contrary-to-fact, especially different-author, LRs. As can be observed in Figure 10(b) and Figure 10(c), the magnitude of the contrary-to-fact different-author LRs is seen to be somewhat bigger than that of the corresponding pre-fusion FTC systems.

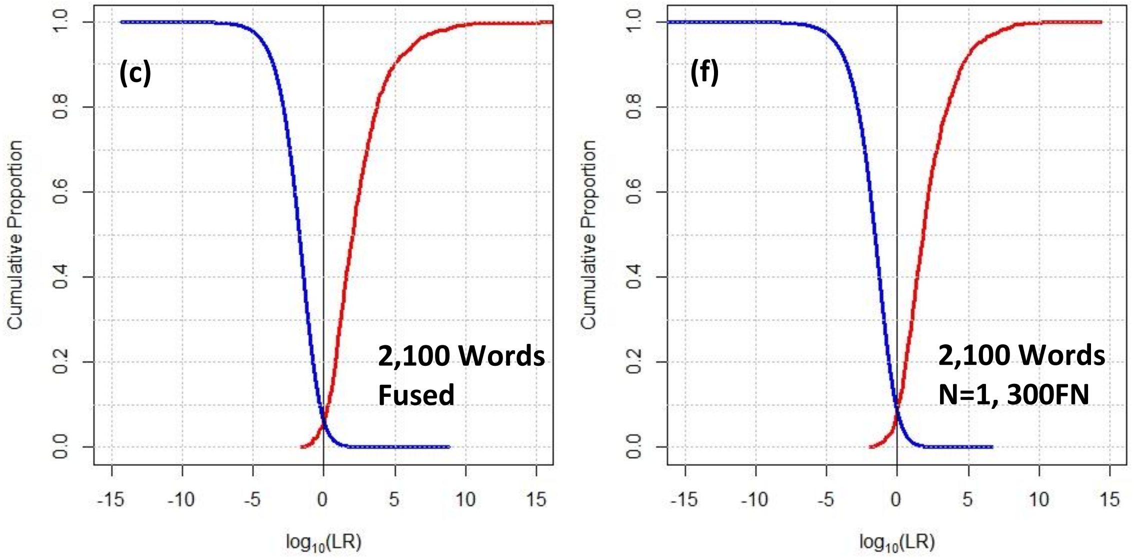
Figure 10

Tippett Plots of Fused System between LRs Derived with Different Word N-gram (n=1,2,3) Features in Comparison to Those of Best-performing Experiment Settings of Word N-gram Features for 700, 1,400, and 2,100 Word Lengths



Note. N stands for the number of word n-gram, while FN stands for the number of features.

Figure 10 (Continued)



Note. The magnitude of the consistent-with-fact different-author LRs (represented in blue solid lines left to the zero threshold) goes beyond the range of x-axis.

Minimising the magnitude of contrary-to- fact LRs, both before and after the LRF, while maintaining, or even improving, the magnitude of consistent-with-fact LRs, is a future study topic that should be pursued.

5.5 Discussion

Although low in dimensional compared to WN2s and WN3s, WN1s have outperformed the other numbers of WNGs. Moreover, it does not take long at all for WN1s (300 WN1s) to achieve their full potential for every document length. The high discriminating power of WN1s in this thesis resonates some observations in the previous author analysis studies. An observation made by Coyotl-Morales et al. (2006) and Sanderson and Guenter (2006) suggested that stylometric features built on a WNG model with the higher numbers of n do not yield an outstanding classification accuracy compared to individual word features or WN1s. This is because the dimensionality of the identification task increases as an n increases to account for the increasing possible WNG combinations. WNGs with the high number of n may result in the sparse representation of WNGs as most WNG combinations will not be found in an WNG-modelled document, especially in a short one (Stamatatos, 2009). Furthermore, Gamon (2004) also pointed out higher numbers of WNGs may unintentionally capture content-specific information instead of individual stylistic information. Meanwhile, Sari et al. (2018) also supported this stance, saying that the content-based features, including WN1s, are highly effective in dealing with authorship analysis tasks on data that have high topic diversity.

With regard to LR-based FTC, I think the observations made in authorship analysis studies also apply to the phenomenon we are experiencing now. One possible explanation is when the documents in comparison are modelled using the WN1 model, the documents are modelled into the strings of the resultant individual word features, or to put it simply, words, with their associated frequencies. When individual words are used as features, they are more likely to be found used in the testing database than WN2s or WN3s. This makes the model representation of WN1s denser, making WN1s able to capture the stylistic information better than the higher numbers of WNGs. According to Sari et al. (2018), WN1s are supposed to work well since product reviews are written in a variety of topics. That said, ambiguity still prevails as to why only 300 features, or in this case words, outperform the higher numbers. I reckon that due to a short amount of data (i.e. only 700 to 2,100 words), the documents in comparison cannot handle the high dimensionality of the FTC task generated by the high numbers (~600) of features (words) as effectively as the lower (~300) numbers. The fluctuation of the FTC system performance as a function of the number of features here may warrant a future study.

Chapter 6 Character N-gram Features: Results and Discussion

The aim of this chapter is to present the results of the FTC system built by using character n-grams (CNGs) to derive same-author and different-author LRs for different experiment setting. I will be presenting the best-performing experiment settings using CNGs in a table and visualising how the FTC system performance changes as a function of feature numbers. The fused FTC system between different CNGs (n=1,2,3,4) will then be presented. Discussion as to the implications of the numerical results and the visualisation of the LRs derived will then be made. For full results of the experiment settings at each number of CNGs, see Appendix C.

6.1 Best-performing Character N-gram Features

Table 7 presents the best-performing experiment settings of CNGs for every document length. Unlike word n-gram features (WNGs), an n-gram of two (n=2) of CNGs, or character bigrams (CN2s), is the best-performer for character n-gram features, securing a place in every experiment setting in Table 7. At its highest capacity, the best-performing experiment settings of CNGs, which are all CN2s, yield the *Cllr* values of 0.722, 0.498, and 0.356, and the $Cllr^{min}$ values of 0.710, 0.483, 0.342, for the documents of 700, 1,400 and 2,100 words respectively. Worth notetaking is that the number of CNGs that performs best is now the maximum number of CNGs features tested, which is 1,000. This is also unlike the experiments on WNGs where the best-performing number of features is only 300 with WN1s.

6.2 Performance of Character N-gram Features as a Function of the Number of Features

The performances of the three different CNGs can be visually observed in Figure 11. In general, when the lower numbers of features (<400 features) are used, the point representing the C_{llr} values of all the numbers of CNGs are closely grouped together as

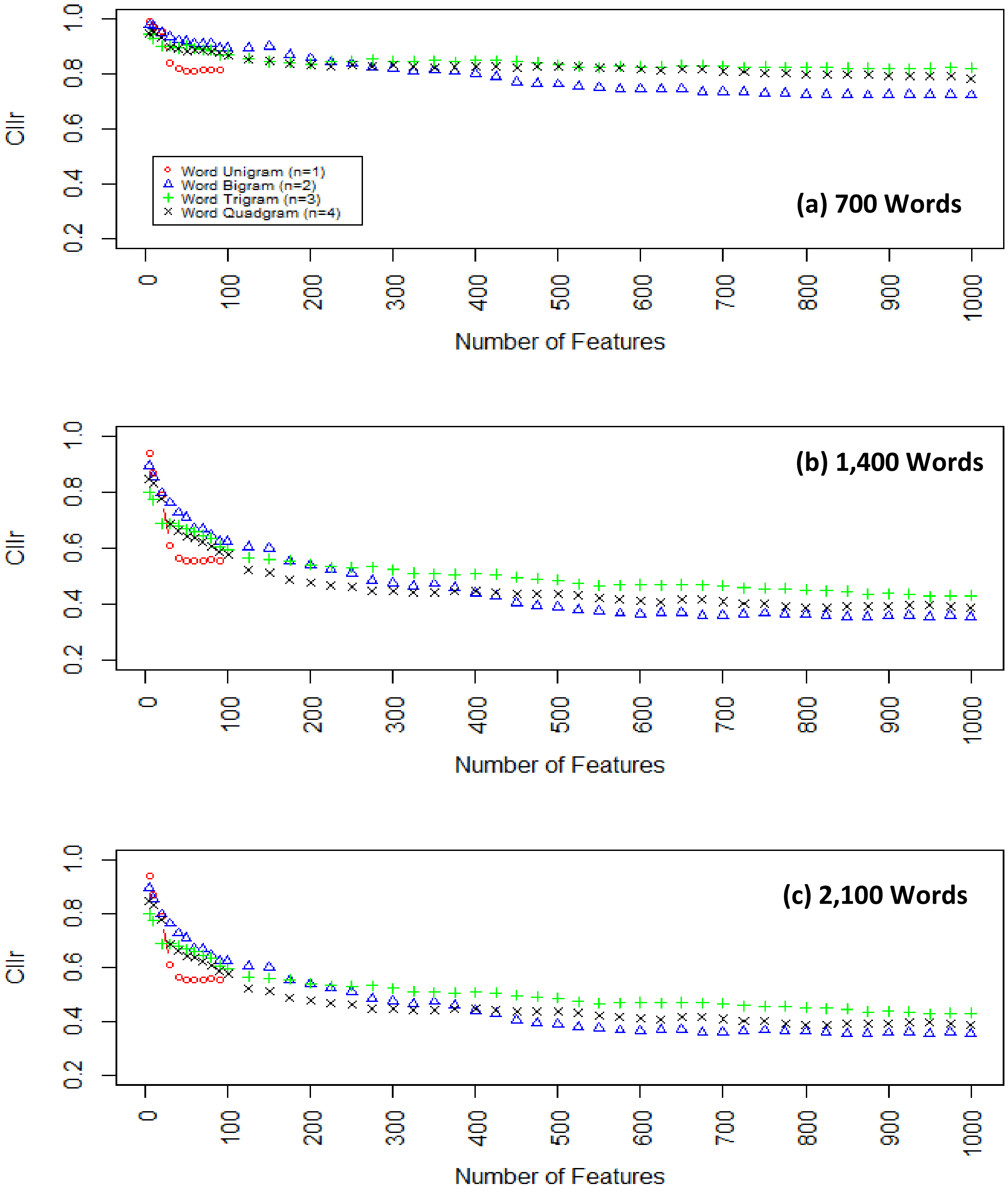
Table 7

Best-performing Experiment Settings of Character N-gram Features for 700, 1,400, and 2,100 Word Lengths

Sample Word Number	Number of N-gram	Number of Features	*Cllr*	$Cllr^{min}$
700	2	1,000	0.722	0.710
	2	875	0.722	0.708
	2	825	0.723	0.710
	2	975	0.724	0.710
	2	850	0.724	0.711
	2	900	0.724	0.711
	2	950	0.725	0.711
	2	925	0.726	0.710
	2	800	0.726	0.713
	2	775	0.729	0.714
1,400	2	1,000	0.498	0.483
	2	875	0.500	0.487
	2	700	0.501	0.487
	2	975	0.501	0.489
	2	900	0.501	0.487
	2	950	0.502	0.488
	2	850	0.502	0.489
	2	675	0.503	0.491
	2	825	0.503	0.489
	2	725	0.504	0.490
2,100	2	1,000	0.356	0.342
	2	875	0.356	0.344
	2	850	0.357	0.344
	2	950	0.357	0.342
	2	900	0.358	0.347
	2	975	0.358	0.343
	2	925	0.358	0.344
	2	825	0.359	0.346
	2	675	0.360	0.349
	2	700	0.362	0.350

Figure 11

C_{llr} Values Plotted against Number of Features of Character Unigrams (n=1), Character Bigrams (n=2), Character Trigrams (n=3), and Character Quadgrams (n=4) for 700, 1,400, and 2,100 Word Lengths



Note. The maximum number of character unigrams (n=1) tested is 90.

Table 8

Results of Fused FTC System Using Character N-gram Features (n=1,2,3,4; Top Table) in Comparison to Results of Best Experiment Settings of Character N-gram Features (Bottom Table) for 700, 1,400, and 2,100 Word Lengths

Sample Word Number	*Cllr*	*Cllr*^{min}
700	0.679	0.666
1,400	0.433	0.421
2,100	0.297	0.228

Sample Word Number	Number of N-gram	Number of Features	*Cllr*	*Cllr*^{min}
700	2	1,000	0.722	0.710
1,400	2	1,000	0.498	0.483
2,100	2	1,000	0.356	0.342

observed in Figure 10, which means that the FTC system performances of different numbers of character n-grams tend not to differ from each other much during the lower numbers of features. However, when there are more than 400 features in the FTC system, the FTC system performances across CN2s, character trigrams (CN3s), and character quadgrams (CN4s) start to part from each other. Like character bigrams, CN3s and CN4s all take very long in term of the number of features to reach their peaks, with CN4 slightly outperforming CN3s (see Figure 10). CN1s, despite being tested only at 90 features maximum (in this case characters) due to the reason mentioned in Section 2.4.2.2, yield surprisingly good results with their best performers providing the C_{llr} values of 0.807, 0.655, and 0.553 for 700, 1,400, and 2,100 words respectively. Interesting is the fact that it does not take long for CN1s to hit their performance ceilings for 700 (50 features) and 1,400 words (60 features); but for the longer documents as in those of 2,100 words, it takes 90 features for CN1s to reach their peak. This resonates with the authorship discrimination performance of WN1s which also needs the mid-range number of n-grams tested (300 out of 600 features) to reach their optimal point.

6.3 Fused FTC System Between Character N-gram Features

When the LRs derived from different CNGs (n=1,2,3,4) are logistic regression fused (LRF), the fused FTC system once again considerably outperforms the separate pre-fusion FTC systems as shown in Table 8. The LRF results in the improvements of the *Cllr* values of 0.043, 0.065, and 0.059, and the improvements of the *Cllr*$^{\mathrm{min}}$ values of 0.044, 0.062, and 0.114, for the documents of 700, 1,400, and 2,100 words respectively.

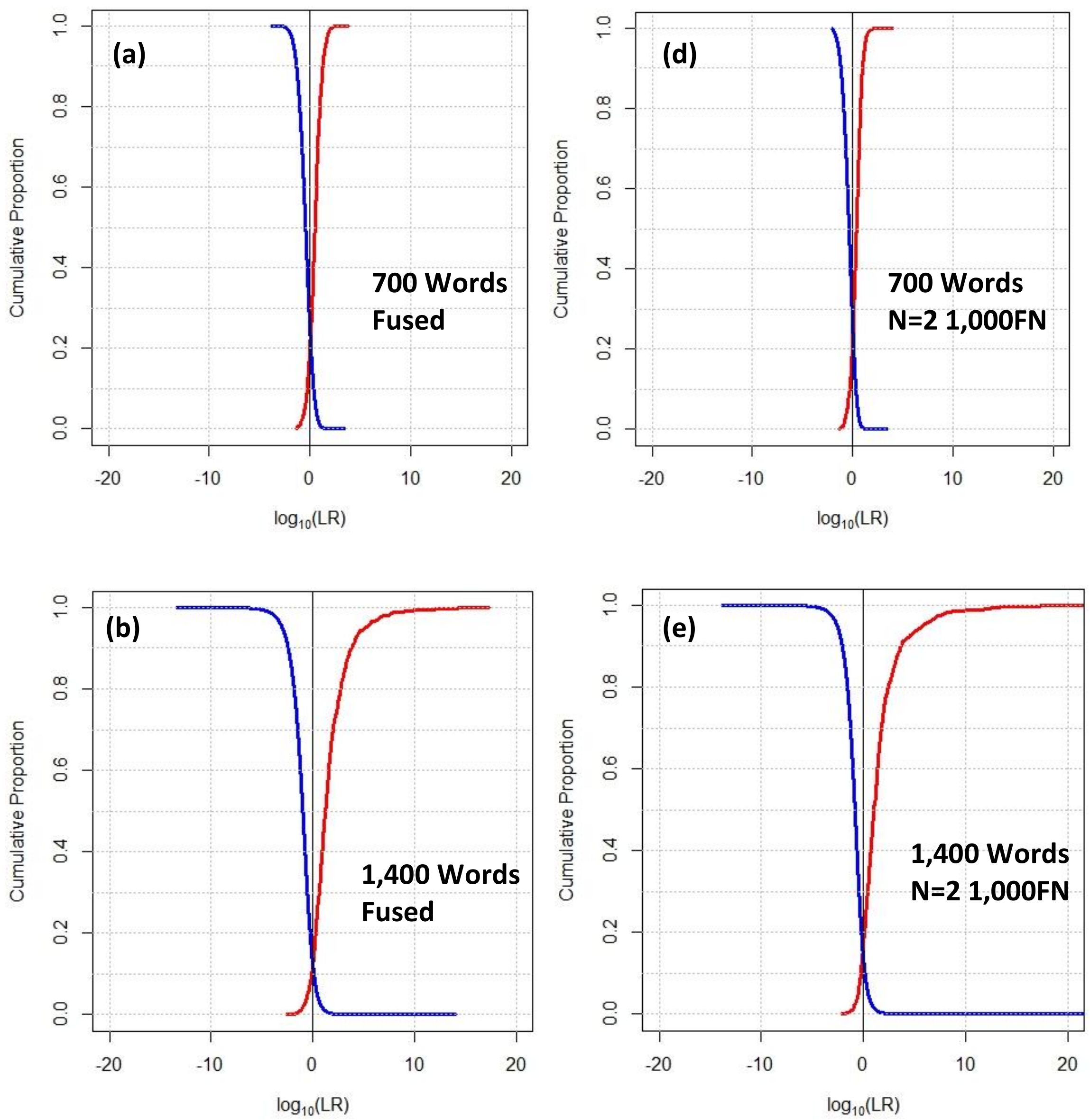
What the fused FTC system of CNGs have in common with that of WNGs is that the LRF seems to work better with documents of longer length (i.e. 1,400 and 2,100 words) as the C_{llr} and *Cllr*$^{\mathrm{min}}$ improvements for the documents of longer length are higher than that of the documents of shorter length for both fused systems.

6.4 Tippett Plots of Fused FTC System Between Character N-gram Features

The magnitude of the LRs derived with the different CNGs, in comparison to that of the corresponding systems, can be visually observed in Figure 12. Although the degree of improvement in terms of C_{llr} is smaller in CNGs than in WNGs after the LRF, the effects of the LRF can still be observed. For instance, visual inspection of Figure 12(b) and Figure 12(e) reveals that for the documents of 1,400 words the LRF has the magnitude of the fused contrary-to-fact different-author LRs smaller than that of the pre-fusion counterparts (i.e. from exceeding the magnitude of 20 to around the magnitude of 13). For the documents of 2,100 words, the LRF also causes the magnitude of the fused contrary-to-fact different-author LRs to be smaller than the unfused counterparts. However, partly because of the lower C_{llr} discrepancies, the Tippett plots for the documents of 700 words are almost identical to each other with only minor improvements in fused consistent-with-fact LRs. In sum, the logistic regression is seen to bring some positive improvement towards the fused FTC system performance, especially in minimising the magnitude of the contrary-to-fact LRs.

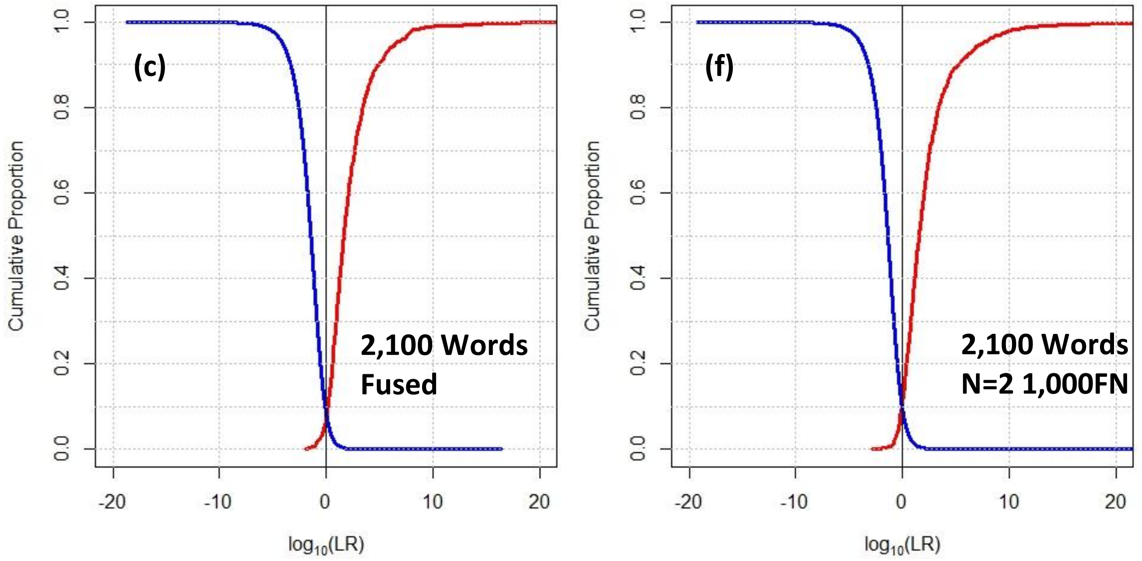
Figure 12

Tippett Plots of Fused System between LRs Derived with Different Character N-gram (n=1,2,3,4) Features in Comparison to Those of Best-performing Experiment Settings of Character N-gram Features for 700, 1,400, and 2,100 Word Lengths



Note. The magnitude of the consistent-with-fact LRs in Figure 11(e) goes beyond the range of x-axis

Figure 12 (Continued)



Note. The magnitude of the consistent-with-fact same-author LRs in Figure 11(c) and Figure 11(f), and the magnitude of the contrary-to-fact different-author LRs goes beyond the range of x-axis.

In terms of fixing errors, the LRF does wonders to the fused LRs derived with different CNGs. However, it cannot be denied that there are still the considerable magnitude of contrary-to-fact LRs to help correct, especially for the documents of 1,400 and 2,100 words. The reduction in the magnitude of contrary-to-fact LRs also comes at the cost of the reduction in the magnitude of consistent-with-fact LRs. For example, considering Figure 12(b) and Figure 12(e), the counterfactual same-author LRs gets much more conservative after the LRF (from exceeding the magnitude of 20 to around the magnitude of 17). Ways to minimise the counterfactual LRs while also maintain or improve the consistent-with-fact LRs are a topic of focus in future research.

6.5 Discussion

The fact that CN2s work best in the FTC system of CNGs here may contribute a surprising finding to authorship analysis studies. As several authorship analysis studies discussed, the higher numbers of CNGs, usually from three to five, yielded best results for documents of medium (~1,000) to long (>10,000) length (Stamatatos, 2009; Peng et al., 2003). For instance, Keselj et al. (2003) found out that the higher numbers of CNGs, including four, five, six, seven, and eight, yielded the best classification accuracies for authorship attribution tasks on their English datasets. Despite the high discriminating power of the higher numbers of CNGs in many authorship analysis studies, several studies also report the high discriminating capacity of WN2s as well. Testing CNGs from the range of two to nine in attribution tasks on a newspaper corpus, Grieve (2007) achieved high classification accuracies for using CN2s. Forsyth and Holmes (1996) conducted a study to find out who the original writer of *the Federalist Papers* was, and found that that CN2s, when used in conjunction with other types of features, can effectively capture both stylistic and content information of the documents in comparison.

What is obvious now is that CN2s may not work as effectively as for this thesis in other author attribution tasks or different experiment settings; however, ambiguity still looms as to the reason why. One possible explanation is that since the previous authorship analysis studies (Stamatatos, 2009; Peng et al., 2003) made use of long documents, accurate CNG models could be built with the high numbers of CNGs. Thus, if this thesis had made use of documents of longer length, CN3s or CN4s might have worked better than CN2s. Another possible reason is the topic diversity of the FTC task conducted for this thesis. Grieve (2007) pointed out that although the higher numbers of CNGs may be a good indicator of topic, they cannot be a good indicator of authorship or style at the same time. Grieve also argued that since his study conducted on a newspaper corpus achieved good classification accuracies with CN2s, the lower numbers of n-grams may be good at discriminating authorship. For this thesis, product reviews of various topics were pooled together to be concatenated into author profiles, so such an FTC task is not a topic-based classification as in Peng et al.(2003) or Keselj et al. (2003). This may be the reason why CN2s work best for the topic-diverse FTC task conducted for this thesis.

Another worth discussing topic is the number of features that yields the optimal results for CNGs. As discussed earlier, 1,000 CNGs work best for every document length. This seems to be the case for many authorship analysis studies irrespective of which number of CNGs works best. For example, Houvadas and Stamatatos (2006) conducted an experiment using the higher numbers of CNGs (from three to five) to identify authorship in a newspaper corpus, and their classification algorithm works best when it is working with the highest number of CNGs. In Grieve (2007), the classification algorithm works best for the lower numbers of CNGs when the highest number of CNGs that has the minimum count of ten is set. Stamatatos (2013) argued for the usefulness of CNGs in dealing with the high dimensional data that since they are smaller than the other types of features (e.g. WNGs and PNGs), each number of CNGs can be better trained with a higher amount of data. According to the previous studies, I suspect that the CNGs would work even better if the maximum number of features went beyond 1,000. For now, based on this thesis, it seems to me that CNGs work best when they are used at the maximum number that a text comparison can manage. However, the optimal number of features may change depending on different conditions of authorship analysis experiments.

Chapter 7 Part-of-speech N-gram Features

The aim of this chapter is to explore the results of the FTC system with part-of-speech n-gram features (PNGs). As with the other results chapters, I will be presenting the best-performing experiment settings of PNGs and visualizing how the FTC system performance changes as a function of feature numbers. After that, the fused FTC system between PNGs and the magnitude of the fused LRs derived with PNGs will be presented. Eventaully, I will discuss the significance of the results. For full results of the experiment settings at each number of PNGs, see Appendix D.

7.1 Best-performing Part-of-speech N-gram Features

Table 9 gives us an overview of the best-performing experiment settings of PNGs for every document length with their *Cllr* and $Cllr^{min}$ values. The best performers among PNGs are part-of-speech bigrams (PN2s) with their appearances in the top-ten best performers for every document length. The numbers of features that perform best here are those that fall in the mid-to-high tested range (out of 600 features), which are 525 for the documents of 700 words and 400 for those of 1,400 and 2,100 words. The FTC system performance with PNGs under the best experiment circumstances yields the *Cllr* values of 0.714, 0.486, and 0.342, and the $Cllr^{min}$ values of 0.696, 0.473, 0.329 for the documents of 700, 1,400, and 2,100 words respectively. What is also worth pointing out is that the documents of shorter length (i.e. 700 words) require the higher number of features (i.e. 525 features) for PNGs to reach their performance potential, while those of longer length (i.e. 1,400 and 2,100 words) need the lower number of features (i.e. 400 features) to do so. A question may arise as to why 525 PN2s work best for the documents of shorter length but not for those of longer length. This is counterintuitive as, in my opinion, the shorter documents should need the lower number of features while the longer documents the higher number of features.

Table 9

Best-performing Experiment Settings of Part-of-speech N-gram Features for 700, 1,400, and 2,100 Word Lengths

Sample Word Number	Number of N-gram	Number of Features	*Cllr*	$Cllr^{min}$
700	2	525	0.714	0.696
	2	450	0.715	0.698
	2	475	0.715	0.700
	2	425	0.717	0.696
	2	500	0.717	0.701
	2	550	0.718	0.700
	2	400	0.719	0.697
	2	600	0.721	0.708
	2	575	0.722	0.705
	2	375	0.725	0.703
1,400	2	400	0.486	0.473
	2	425	0.487	0.474
	2	525	0.488	0.474
	2	600	0.488	0.476
	2	500	0.489	0.477
	2	450	0.489	0.478
	2	475	0.489	0.476
	2	550	0.489	0.476
	2	375	0.490	0.478
	2	575	0.491	0.479
2,100	2	400	0.342	0.329
	2	425	0.343	0.331
	2	525	0.344	0.331
	2	475	0.345	0.332
	2	450	0.347	0.334
	2	500	0.348	0.334
	2	550	0.350	0.335
	2	600	0.350	0.336
	2	375	0.351	0.340
	2	575	0.355	0.340

7.2 Performance Saturation of Part-of-speech N-gram Features

The comparison of the performances of different PNGs is visually represented in Figure 13. The performances of all the numbers of PNGs seem to work in a similar tendency; that is, they tend to significantly improve until approximately 50 to 75 features and then hardly improve or deteriorate at all. Considering Figure 13, after the three FTC subsystems have around 50 to 75 features, the points plotted for the C_{llr} values hardly fluctuate at all. This results in that the C_{llr} values of the performance peaks of the three different PNGs for every document length are not much higher than those that are around the peaks.

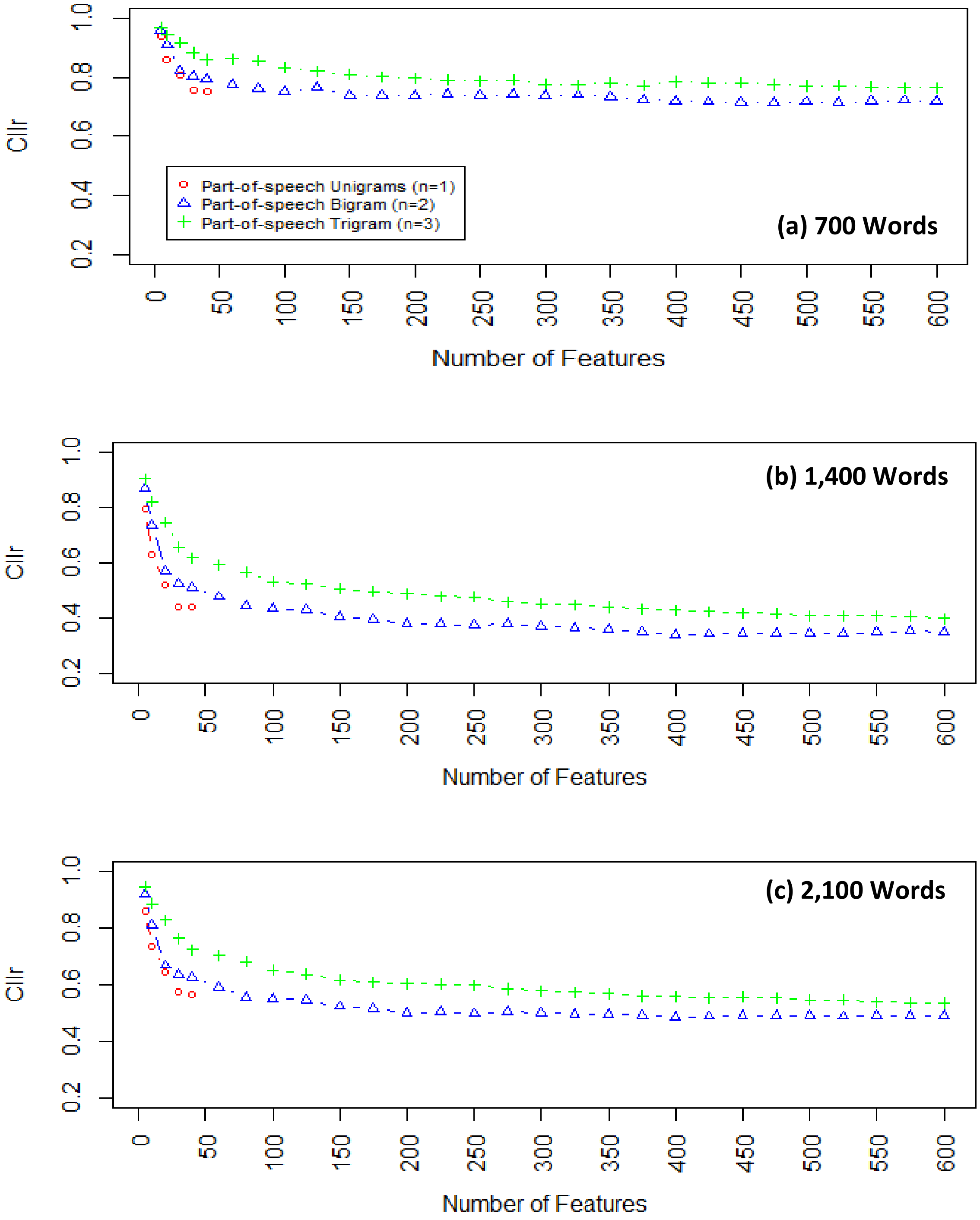
Although being tested at only 40 features due to that part-of-speech categories are far less than words or characters, PN1s put in a competitive performance, with their top-performers boasting the C_{llr} values of 0.752, 0.564, and 0.439 for the documents of 700, 1,400, and 2,100 words respectively. PN1s seem to work best at the higher numbers of features (30 to 40 or part-of-speech tags). PN3s at their best experiment settings are not losing to PN2s by much, with only little C_{llr} discrepancies when compared to the top performers (PN2s) as PN3s claim the C_{llr} values of 0.764, 0.535, and 0.399 for the documents of 700, 1,400 and 2,100 words respectively. Note that for the documents of shorter length (i.e. 700 words), PN1s (C_{llr}=0.752) perform better than PN3s (C_{llr}=0.764).

7.3 Fused FTC System between Part-of-speech N-gram Features

The logistic regression fusion (LRF) was applied to the LRs derived with the different PNGs (n=1,2,3), and the *Cllr* and C_{llr}^{min} values of the fused FTC system with the different PNGs and of the pre-fusion FTC system of PNGs can be observed in Table 10. The fused FTC system performs better than the pre-fusion FTC systems. The fused FTC system of PNGs reports the C_{llr} improvements of 0.036, 0.033, and 0.021 and the C_{llr}^{min} improvements of 0.036, 0.046, and 0.032 for the documents of 700, 1,400, and 2,100 words respectively. Another phenomenon that is unlike what we saw in the previous experiments that the LRF seems not to work with the documents of longer length (i.e. 1,400 and 2,100 words) than those of shorter length anymore as the C_{llr} improvements for the documents of 700 words is the highest (0.036) compared to 1,400 (0.033) and 2,100 (0.021) respectively.

Figure 13

C_{llr} Values Plotted against Number of Features of Part-of-speech Unigrams (n=1), Part-of-speech Bigrams (n=2), and Part-of-speech Trigrams (n=3) for 700, 1,400, and 2,100 Word Lengths



Note. Part-of-speech unigrams (PN1s) are tested at only 40 features maximum.

Table 10

Results of Fused FTC System Using Part-of-speech N-gram Features (n=1,2,3; Top Table) in Comparison to Results of Best Experiment Settings of Part-of-speech N-gram Features (Bottom Table) for 700, 1,400, and 2,100 Word Lengths

Sample Word Number	*Cllr*	*Cllr^{min}*
700	0.678	0.660
1,400	0.453	0.440
2,100	0.321	0.310

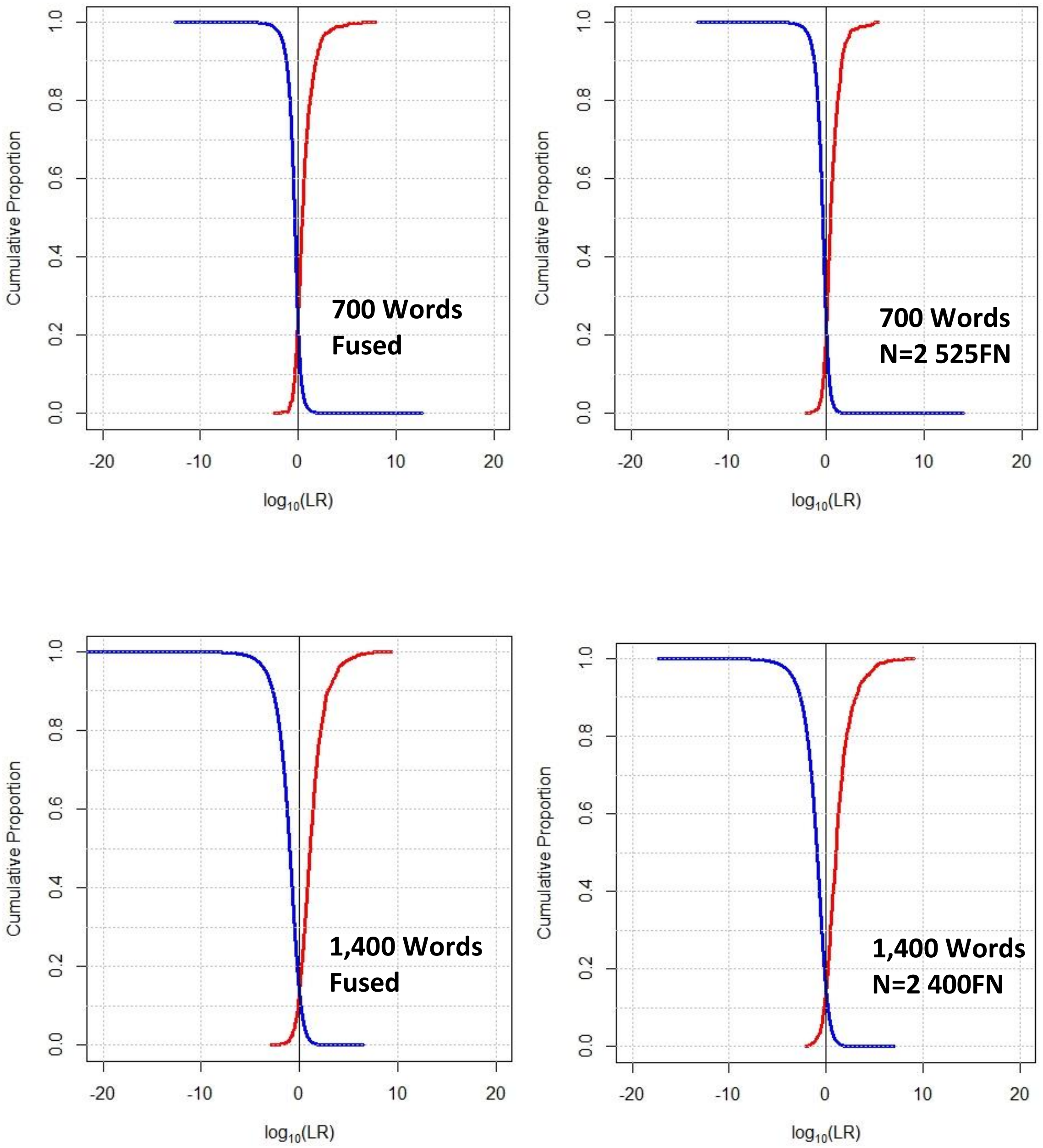
Sample Word Number	Number of N-gram	Number of Features	*Cllr*	*Cllr^{min}*
700	2	525	0.714	0.696
1,400	2	400	0.486	0.473
2,100	2	400	0.342	0.329

7.4 Tippett Plots of Fused FTC System Between Part-of-speech N-grams

Figure 14 displays the Tippett plots of the fused FTC system between the different PNGs compared to those of the pre-fusion FTC system of PNGs. The performance improvement in terms of *Cllr* between pre- and post-fusion in not extensive for PNGs; accordingly, the magnitude of improvement manifested in the Tippett plots of the fused LRs is not as evident. Considering Figure 14(a) and Figure 14(d), even the fused and pre-fusion FTC systems for the documents of 700 words, which have the highest C_{llr} improvement (0.036) out of three, only see some minimal improvements to the consistent-with-fact different-author LRs and the contrary-to-fact same-author LRs. Unlike the document of 700 words, the effect of the LRF is more noticeable for the documents of longer lengths (i.e. 1,400 and 2,100 words). For the document of 1,400 words, the magnitude of the fused consistent-with-fact different-author LRs of extends to the magnitude of Log10LR of -20. For the documents of 2,100 words where the C_{llr} improvement is lowest (0.021), there are

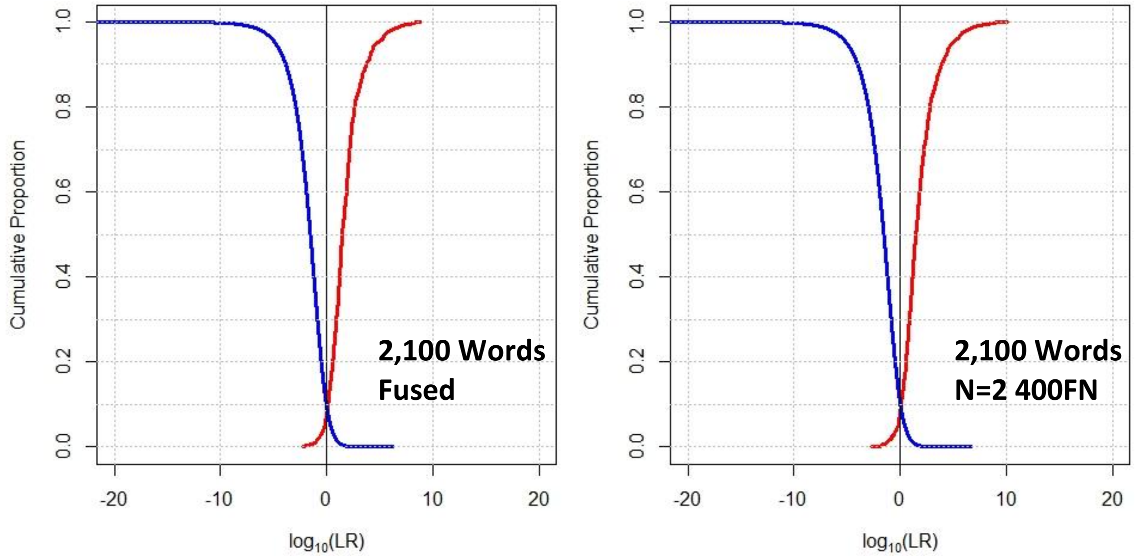
Figure 14

Tippett Plots of Fused System between LRs Derived with Different Part-of-speech N-gram (n=1,2,3) Features in Comparison to Those of Best-performing Experiment Settings of Part-of-speech N-gram Features for 700, 1,400, and 2,100 Word Lengths



Note. The magnitude of the consistent-with-fact different-author LRs in Figure 13(b) goes beyond the range of x-axis.

Figure 14 (Continued)



Note. The magnitude of the consistent-with-fact different-author LRs in Figure 13(c) and Figure13(f) goes beyond the range of x-axis.

barely any visual differences between the Tippett plots of the fused and pre-fusion FTC system. Having been minimised to a small extent, the magnitude of contrary-to-fact LRs for both fused and pre-fusion FTC systems are still noticeable. This again would be a subject of future study to see to these faults.

7.5 Discussion

Unlike the previous studies, it is interesting to see the superior performance of PN2s in the current study. PN2s can capture sequential information, something that PN1s cannot comprehend. Between PN2s and PN3s, the occurrences of the latter are naturally less than PN1s and PN2s; this naturally makes PN3s less supported by data. Therefore, PN2s can be seen to be a perfect balance of PN1s and PN3s: they can capture hidden individual authorial information of text and allow the reoccurrences of PNGs as they are not as high dimensional as PN3s.

Different performance saturation points among PNGs warrants some discussion. For every number of PNGs, there is in need of higher numbers of features to achieve the optimal results. However, PNGs does not need to reach the maximum number of features tested in this thesis to reach their optimal points. I believe that this has to do with nature of PNGs. That said, for this thesis, the high, but not maximum, numbers of PNGs may be enough to best capture stylistic information, as the individual features themselves (words/POS tags or a string of words/POS tags) can at least tell us something without complementarities with the other features in their respective types of features. Unlike CNGs which are considered more statistically auspicious than the other types of n-gram features, they need the highest number of features allowed to be able to have the same level of the discriminating power as the other types of features because the individual features themselves (characters or a string of characters) do not tell much as to individual writing style. The FTC system performance of PNGs, especially of PN2s and PN3s, seems to be stable, or deteriorating by a slight margin, after they reach their optimal points. This is probably because when the number of features increases, there tends to be the problem of data sparsity which causes a lot of zero occurrences of PNGs of higher feature numbers for some documents of some authors, forcing the FTC system performance not be able to gain new information more than those features of lower feature numbers.

Chapter 8 Fused FTC System between Statistical, Word N-gram, Character N-gram, and Part-of-speech N-gram Features

The aim of this chapter is to present and discuss the results of the fused FTC system between the different types of features tested in this thesis. Firstly, I will be presenting and discussing the fused FTC system performance of each feature type and compare it to one another. Eventually, we will be investigating and discussing the fused FTC system performance between all the types of features to see how much it is improved from the fused FTC systems of each feature type.

8.1 Fused FTC System Performance of All Feature Types

Table 11 displays the performances of the best-performing combinations of statistical features (SFs; Section 4.1), and the fused FTC systems of word n-grams (WNGs; Section 5.3), character n-grams (CNGs; Section 6.3), and part-of-speech n-grams (PNGs; Section 7.3). As is evident, comparing the different fused systems of the different feature types, WNGs yield the best fused results, claiming the *Cllr* values of 0.665, 0.363, and 0.230, and the *Cllr*^{min} values of 0.651, 0.349, and 0.219 for the documents of 700, 1,400 and 2,100 words respectively as observed in Table 11(b). The second-to-best performing fused FTC system is that between the different character n-gram features (Table 11(c)), while the second-to-last performing fused FTC system is that between the different part-of-speech n-gram features (Table 11(d)). Note that for the documents of shorter length (i.e. 700 words), the fused FTC system between PNGs works slightly better than that between CNGs, as can be seen in the *Cllr* discrepancy of 0.001 and the *Cllr*^{min} discrepancy of 0.006 between the two fused FTC systems. The best-performing combinations of SFs (Table 11(a)) generally perform worse than the three fused FTC systems; their *Cllr* values differ to those of the best-performing fused FTC system (that of WNGs) by margins of 0.149, 0.309, and 0.364, and

Table 11

Performances of Fused FTC Systems of Statistical Features (a), Word N-gram Features (b), Character N-gram Features (c), and Part-of-speech N-gram Features (d) for 700, 1,400, and 2,100 Word Lengths with the Highest Cllr and Cllr^{min} values Highlighted in Bold Separately for Each Word Length

(a) Statistical Features

Sample Word Number	Combination of features	*Cllr*	*Cllr^{min}*
700	(3,4,6,7,8,9)	0.814	0.792
1,400	(1,2,4,5,6,7,8,9)	0.669	0.652
2,100	(1,4,6,7,8,9)	0.594	0.571

(b) Word N-gram Features

Sample Word Number	*Cllr*	*Cllr^{min}*
700	**0.665**	**0.651**
1,400	**0.363**	**0.349**
2,100	**0.230**	**0.219**

(c) Character N-gram Features

Sample Word Number	*Cllr*	*Cllr^{min}*
700	0.679	0.666
1,400	0.433	0.421
2,100	0.297	0.228

(d) Part-of-speech N-gram Features

Sample Word Number	*Cllr*	*Cllr^{min}*
700	0.678	0.660
1,400	0.453	0.440
2,100	0.321	0.310

Note. (a) is the best-performing combinations of statistical features for every word length (700, 1,400, and 2,100 words). (a) is adapted from Table 3 (Section 4.1), (b) Table 6 (Section 5.3), (c) Table 8 (Section 6.3), and (d) Table 10 Section (7.3).

Table 12

Performances of Fused FTC Systems All Types of Features Including Statistical Features, Word N-gram Features, Character N-gram Features, and Part-of-speech N-gram Features for 700, 1,400, and 2,100 Word Lengths

Sample Word Number	*Cllr*	*Cllr^{min}*
700	0.569	0.554
1,400	0.309	0.298
2,100	0.192	0.183

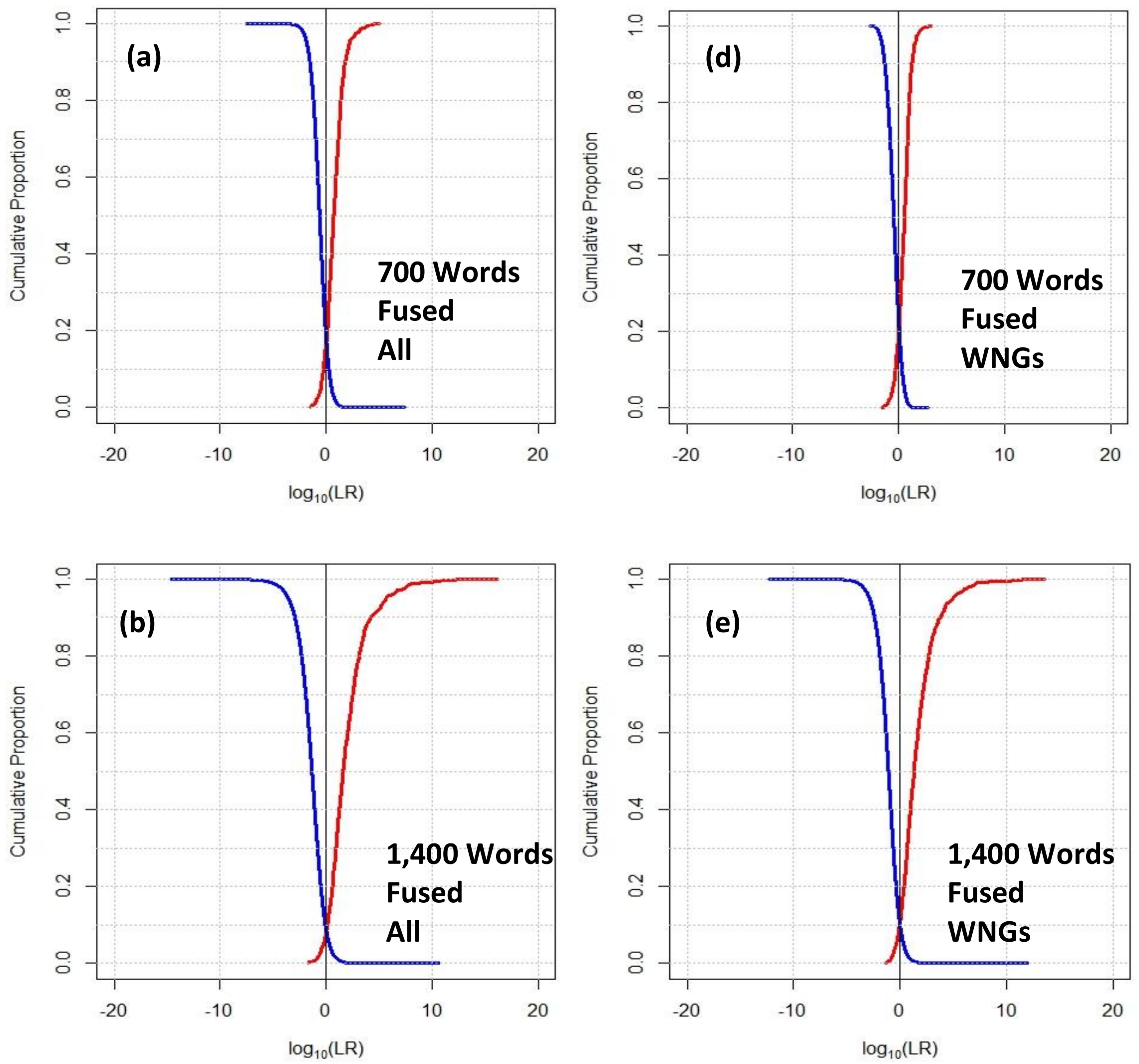
their *Cllr*^{min} values by margins of 0.141, 0.303, and 0.352, for the documents of 700, 1,400, and 2,100 words respectively.

8.2 Fused FTC System Performance between All Types of Features

Table 12 displays *Cllr* and *Cllr*^{min} values of all the types of features tested for thesis combined, which are SFs, WNGs, CNGs, and PNGs. In comparison to the best-performing fused FTC system of a feature type, which is that of WNGS as mentioned in Section 8.1, the fused FTC system between all the feature types enjoys the *Cllr* improvements of 0.096, 0.054, and 0.042, and the *Cllr*^{min} improvements of 0.097, 0.051, and 0.036, for the documents of 700, 1,400, and 2,100 words respectively. Evident is that the improvements in the fused FTC system performance between all the feature types fluctuate with the document length. It can be observed that the LRF obviously improves the fused FTC system performance between all the feature types, but as the document length increases, the *Cllr* and *Cllr*^{min} improvements gradually weaken.

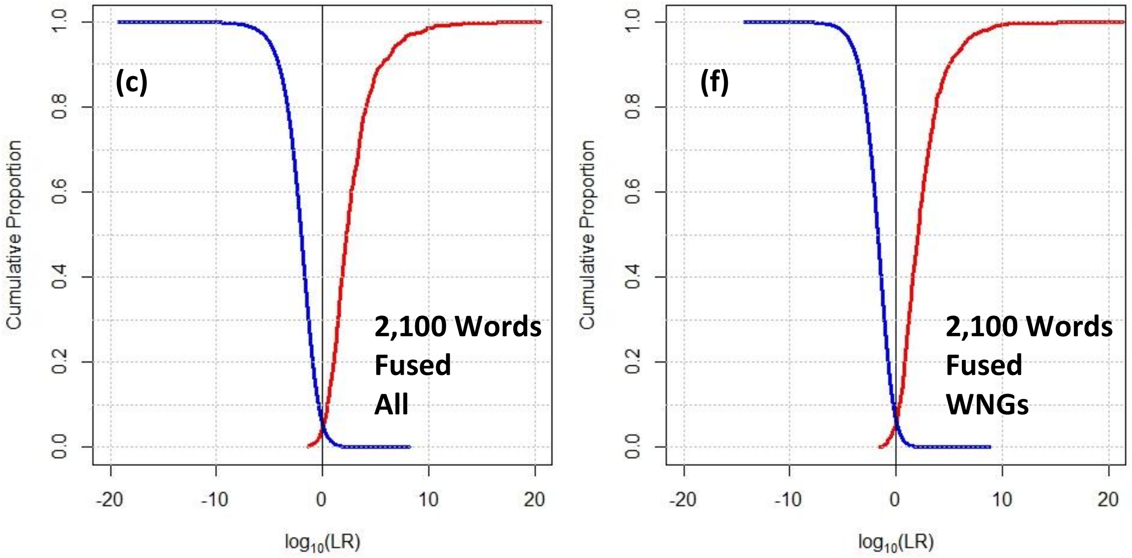
Figure 15

Tippett Plots of Fused System between LRs Derived with Different Types of Features Including Statistical, Word N-gram (n=1,2,3), Character N-gram (n=1,2,3,4), and Part-of-speech N-gram (n=1,2,3) Features in Comparison to Those Derived with Different Word N-gram Features for 700, 1,400, and 2,100 Word Lengths



Note. ‘All’ indicates that the Tippett plots show the LRs derived with all the feature types, while ‘WNGs’ indicates that the Tippett plots show the LRs derived with fused word n-gram features. The magnitude of the consistent-with-fact same-author LRs goes beyond the range of x-axis in Figure 14(f).

Figure 15 (Continued)



8.3 Tippett Plots of Fused FTC System Performance between All Types of Features

Figure 15 visualises the quality of the derived LRs of the fused FTC system performance between all the features in comparison to those of the fused FTC system performance between the different WNGs, the best-performing individual fused FTC system.

In comparison to the Tippett plots showing the fused LRs derived with the different WNGs, the magnitude of the fused LRs derived with all the types of features has seen some improvements. For example, for the documents of 700 words, the magnitude of the fused consistent-with-fact same-author and different-author LRs derived with all the types of features in Figure 15(a) goes farther away from the zero threshold than that with the different WNGs in Figure 15(d). For the documents of 1,400 words, the fused consistent-with-fact LRs, both same-author and different-author, derived with all the types of features are also generally greater, as in Figure 15(b), than that of the fused consistent-with-fact LRs derived with the different WNGs, as in Figure 15(e). The LRF also helps minimise the magnitude of the fused contrary-to-fact LRs; as for the documents of 2,100 words, the magnitude of the fused contrary-to-fact different-author LRs derived with all the types of features, as in Figure 15(c), is weaker than that with the different WNGs, as in Figure 15(f). All the improvements in the magnitude of the fused LRs described above contribute to improved values for the metrics (as reflected in the *Cllr* improvements described in Section 8.2) and the better discriminating power (as reflected in the *Cllr*^{min} improvements in Section 8.2) than the fused FTC system with the different WNGs.

Despite some improvements, the LRF also brings about some downgrades to the fused FTC system performance between all the types of features. The most obvious is the greater fused contrary-to-fact different-author LRs for the documents of 700 words (from the magnitude of almost 5 in Figure 15(d) to almost 8 in Figure 15(a)). The magnitude of The LRF also weakens the fused consistent-with-fact LRs in some cases; as for the documents of 2,100 words, it weakens the magnitude of the fused consistent-with-fact same-author LRs (from the magnitude of exceeding 20 in Figure 15(f) to approximately 20 in Figure 15(c). As I have been constantly saying throughout this thesis, future research that looks into how to improve the LRF performance in LR-based FTC or techniques that can minimise the magnitude of the contrary-to-fact LRs while maintain, or in a better case, improve, the magnitude of the consistent-with-fact LRs.

8.4 Discussion

8.4.1 Observation about Fused FTC System Performance of Each Type of Features

The best performance of the fused FTC system between WNGs in this thesis has not been foretold by existing literature. To begin with, in authorship analysis studies that experimented on various types of features, the complementarities of WNGs, when two or more numbers of WNGs are combined, have often been regarded as not as useful as those of the other types of features and therefore should be used in conjunction with other types of features (Gamon, 2004; Sari et al, 2018). For example, Sari et al. (2018) conducted an authorship attribution task on various datasets (e.g. judgment writings, newspapers, and movie reviews, etc.), featuring the combination of WNGs, the combination of CNGs, and the combination of both in their classification algorithm; they found that the combination of CNGs and the combination of WNGs and CNGs are generally superior to that of only WNGs in terms of the classification accuracies across all the datasets. It is reported that the lesser contribution of WNGs to a classification algorithm as a whole may result from that WNGs, especially of higher numbers, may not be found at all in the testing database, especially for short documents, forcing bits of stylistic information to go undetected (Stamatatos, 2009, Tambouli & Prasad, 2019).

What I can say now is that the best performance of the fused FTC system between WNGs in this thesis results from topic diversity of the actual database. As discussed in Section 5.5, the database I am working with now had been built by concatenating the authors' product reviews into the author profiles regardless of topics they were writing on; this leads to the database being topic-diverse. As Sari et al. (2018) suggested, content-based features, including WNGs, tend to work better with data that vary in topic. We saw in Section 5.1 that word unigrams (WN1s) work best in the FTC system of WNGs because of how frequent they are found in the testing database regardless of the document lengths. That said, there must be some stylistic or contextual information that WN1s cannot detect, but the LRF would allow word bigrams (WN2s) and word trigrams (WN3s) to fulfill these gaps while also considering the correlations between them. Therefore, when the LRs derived with the different WNGs are fused, the complementarities of those WNGs are bonded, which in turn lead to the better *Cllr* and *Cllr*$^{\text{min}}$ values.

The same reasoning may also be applied to the fused FTC system performances of CNGs and PNGs. Such system performances are not significantly lower than that of WNGs as observed in Table 11, and that probably results from the nature of CNGs and PNGs. In authorship analysis, CNGs are considered features that are efficient in detecting both style and content, while PNGs features that excel in detecting only style (Diederich et al., 2003; Koppel et al., 2011, 2013; Stamatatos, 2013). Working with the actual database that is rich in content, the fused CNGs are able to yield the better fused FTC system performance than PNGs. That said, the fused FTC can still give us new stylistic information as reflected in that the fused FTC system between PNGs is not significantly lower than those of WNGs and CNGs. This may result from the strong complementarities between different stylistic information extracted from the different PNGs.

Another worth notetaking finding is the worst FTC system performance of the best-performing combinations of SFs. Several authorship analysis studies claim that features that are based on word or character statistics are efficient in detecting writing idiosyncrasies in less topic-diverse data more than in more topic-diverse data (Sari et al., 2018; Stamatatos, 2009; Tambouli & Prasad 2019). This is supported by some LR-based FTC studies conducted on the databases that were not as much topic-diverse as the actual database I am working with now. Ishihara (2017a, 2017b) conducted an FTC experiment using the best-performing combinations of SFs on chatlog messages, and under the best experiment setting, Ishihara obtained the *Cllr* value of 0.217 for the documents of 2,500 words. My speculation is that for such a topic-diverse database as the one used in this thesis, the fused SFs may not create strong bonds among all the ten SFs, resulting in the lowest *Cllr* and $Cllr^{min}$ values as seen in Table 11. This also tells us that there may be few to no features that excel in every kind of text; the performance of features highly depends on the quality, quantity, nature, and type of text on which an FTC system is working.

8.4.2 Observation about Fused FTC System Performance between All Types of Features

The improvements in the fused FTC system performance between all the types of features are what we had expected as a successful result of the LRF. However, as mentioned earlier in Section 8.2, the different degrees of the improvements in the fused FTC system performance between all the types of features also warrant some discussion. That is, when the document length increases, the degree of improvement in the fused FTC system performance tends to be less, meaning that the LRF works better for the documents of shorter length than for those of longer length (Ishihara, 2017b, 2021). There is no definitive answer as to this issue now. Further research endeavours that probe into how the performance of the LRF fluctuates as a function of the amount of data may help us answer this big question.

With reference to the other LR-based FTC studies that included maximalising an FTC system performance with the LRF, the degrees of the improvements in this thesis are generally greater. In reference to Ishihara (2017b), the fused FTC system between the three types of features (SFs, WNGs, and CNGs) sees the *Cllr* improvements of 0.014, 0.011, 0.020, and 0.01, and the $Cllr^{min}$ of 0.020, 0.013, 0.018, 0.005, for the documents of 500, 1,000, 1,500, and 2,500 words respectively. In comparison to Ishihara (2021), the fused FTC system between the three distance measures (Euclidean, Manhattan, and Cosine distance measures) gains the *Cllr* improvements of 0.039, 0.076, and 0.072, and the $Cllr^{min}$ improvements of 0.037, 0.075, and 0.070, for the documents of 700, 1,400, and 2,100 words respectively. One main possible reason is that the features used in these LR-based FTC studies are not as extensive as in this thesis, so the improvements in terms of *Cllr* and $Cllr^{min}$ are not as great as those in this thesis.

Chapter 9 Conclusion

Now we are approaching the end of this thesis. What I will firstly be doing now is to revisit the research questions posited in Section 1.4 and explicitly answered them. Then I will evaluate how findings contribute to the further improvement of LR-based FTC. Limitations of this thesis will also be discussed, and future research endeavours that could help answer the unanswered questions in this thesis will be suggested.

9.1 Research Questions Revisited

Along this thesis, the research questions posited in Section 1.4 have already been answered question by question. That said, I will take advantage of this section to explicitly answer all the research questions as follows:

9.1.1 Answering RQ1: 'How well does a specific feature within a feature type perform in the LR-based FTC system?'

As far as we have seen, a specific feature type usually has a specific feature(s) that outperforms the others. As in Section 4.1, findings have already suggested that the best performing SFs are the Flesch-Kincaid readability scores, the ratio of punctuations, the ratio of uppercase characters, the ratio of special characters, and the average number of characters per word. These best-performing SFs, when they are working together, tend to yield the best FTC system performance among the other combinations of SFs. The other well-performing SFs are the vocabulary richness measures such as the type-token ratio (TTR) and Yule's K and the ratio of digits; these SFs can yield good results if they are used with the five best-performing SFs. The worst-performing features are hapax legomenon (i.e. words that appear only once) and the frequency of unusual words.

Section 5.1 probes into the best-performing WNGs. WN1s generally work better than word bigrams (WN2s) and word trigrams (WN3s).

Section 6.1 and Section 6.2 reveal that the best-performing CNGs are CN2s as they generally yield the better *Cllr* values than CN1s and CN3s (see Figure 10 in Section 6.2).

Findings in Section 7.1 and Section 7.2 suggest that PN2s generally yield better performance than PN1s and PN3s

9.1.2 Answering RQ2: 'How well does a specific feature type perform in the LR-based FTC system?'

As presented and discussed in Chapter 8, when all the features within the specific types are fused, the fused FTC systems lead to the better performance than the pre-fusion FTC systems irrespective of which feature type we are talking about (for full results on the LRF of each feature type, see Section 4.1 for SFs, Section 5.3 for WNGs, Section 6.3 for CNGs, and Section 7.3 for PNGs). As presented in Section 8.1, the best-performing fused FTC system is that between the different WNGs. The second-best fused FTC systems are those with the different CNGs and with the different PNGs; these two fused FTC systems yield competitive results, with that of CNGs performing slightly well than that of PNGs. The best-performing combinations of SFs perform worse than the three fused FTC systems. When all the fused FTC systems are fused together, the fused FTC system between all the types of features even yields a greater system performance.

9.1.3 Answering RQ3: 'What is the optimal number of stylometric features for each feature set or experiment setting in the FTC system designed for this thesis?'

Each feature set (for SFs) or experiment setting (for WNGs, CNGs, and PNGs) requires different numbers of features to reach their performance ceilings. The optimal combination number of SFs in each feature set depends on the document length. Only five to six well-performing SFs can yield optimal results for the documents of 700 words, while the document of longer length (i.e. 1,400 and 2.100 words) may need a combination of robust seven to eight SFs to achieve best results (see Section 4.2). WNGs, on the other hand, need

300 features, out of the maximum of 600 features per experiment setting, to reach their best discriminating capacity for every document length (see Section 5.1 and Section 5.2). Unlike WNGs which need only the middle-range number of features (300 out of 600 features), CNGs require the maximum number of features allowed in this thesis (1,000 features) to hit their peaks for every document length (see Section 6.1 and Section 6.2). It is more complicated for PNGs as the different word lengths needs different numbers of features to yield the best results; the documents of shorter length (i.e. 700 words) need 525 features while those of longer length (i.e. 1,400 and 2,100 words) need only 400 features.

9.1.4 Answering RQ4: 'Does the FTC system need further optimization, namely logistic regression calibration?'

The honest answer of RQ4 is no. In Chapter 3, I have already demonstrated that the LRC does not need any significant improvements to the LRs derived via the score-to-LR conversion model. The reason is that the FTC system that has been calibrated once by such a model are well-calibrated already, and the LRs derived are ready to be interpreted as the weight of evidence (see Chapter 3).

9.2 Thesis Evaluation

Overall, this thesis has filled some of the research gaps mentioned in Section 1.3. As there have been a number of features that have yet to be tested in LR-based FTC, it is always wise to try new features and see which feature works well and which feature does not. This thesis is the first of its kind to introduce some new features to LR-based FTC, which are for example the Flesch-Kincaid readability score and PNGs (n=1,2,3). Findings suggest that the newly introduced features work well in LR-based FTC within the similarity-scores approaches. Some features that worked well in the previous LR-based FTC studies conducted using the feature-based and similarity-typicality score-based approached by Ishihara (2014, 2017a, 2017b), have been proven not to be as effective for the FTC task in this thesis, for instance the vocabulary richness measures (TTR and Yule’s K) and the ratio of digits. Apart from the different approaches used in the previous LR-based FTC studies and in my thesis, another possible explanation is that the previous LR-based FTC studies conducted FTC experiments on the database of chatlog messages, while this thesis does so on the database of product reviews. Further optimisation—especially the logistic regression calibration (LRC)—of an FTC system is also explored in this thesis. While the previous FTC experiments demonstrated a mixed reception towards the LRC, as discussed in Chapter 3, the LRC is not needed for the FTC system in this thesis. This is mainly because of the different approaches for calculating LRs in the previous LR-based FTC studies and in my thesis. The similarity-only score-based approach converts the multidimensional feature space to a univariate score space. The univariate score space can be appropriately trained and modelled even with a limited amount of data, resulting in that the score-to-LR conversion model, trained by 719 same-author scores and 516,242 different-author scores obtained from the background database, yielded well-calibrated LRs.

It is worth noting that this LR-based FTC thesis is highly context-dependent. That is, there may be no such things as the best-performing features and the most optimal numbers of features that excel in every FTC task regardless of the type of text and the amount of data. As I have shown throughout this thesis, within a specific type of feature, the FTC system performance fluctuates as a function of the number of features and the amount of data put to test at one experiment setting. What I am saying now is that in LR-based FTC, there may be a need to properly select and empirically decide the types of features and the numbers of features every time there is an FTC task, so the experiment conditions can suit the nature of that FTC task and can, to a certain extent, guarantee that they can yield the most optimal outcomes for that FTC task. This is beneficial for FTC casework since there is no one that can predict the nature of an upcoming FTC task. Nonetheless, this has been made one-step easier by findings of the FTC system specifically designed for this thesis. Future research endeavours to improve LR-based FTC can also ensure that there will be enough resources available for empirically deciding testing conditions for various FTC tasks.

9.3 Limitations

As mentioned in Section 1.1 and Section 2.1, this thesis is solely focused on experimenting an FTC task on one-to-many communication, the mode of communication where one communication sender communicates with multiple communication recipients (e.g. reviews, advertisement, social media posts, etc.). As you would imagine, these two modes are highly different in their nature. Consequentially, what has been shown to discriminate writings well in one communication mode may not do well in another. Furthermore, other platforms of one-to-many communication apart from product reviews also need to be tested in LR-based FTC; it is my suspicion that the set of features will work in other mediums as effectively as in product reviews. Although LR-based FTC on one-to-one communication has already been trailed (Ishihara, 2014, 2017a, 2017b), the features tested and the other procedural details in those previous studies and this thesis are not the same, making it difficult to let the previous studies judge the efficiency of LR-based FTC on one-to-one communication. Only this thesis and few LR-based FTC studies, including Carne and Ishihara (2020) and Ishihara (2021) that have experimented an FTC task on a mode of one-to-many communication, allowing future research opportunities to be undertaken in other modes of one-to-many communication. Therefore, future research attempts aimed at addressing this problem by using the same set of features as trialed in this thesis would be highly beneficial to LR-based FTC studies.

Another worth mentioning limitation of this thesis concerns the availability of data in typical FTC scenarios. As you may have noticed, in order to work at its full potential, LR-based FTC needs a considerable amount of data to partition them into three separate databases of test, background, and development (in case that the LR-based FTC system needs a second calibration or the LRF, see Section 2.2). While this thesis takes advantage of an available online corpus to do so, typical FTC casework sometimes cannot afford the luxury of having well categorised and populated pool of authors and such a large amount of text data. Currently, there are no available databases that consist of text data that have been used in a forensic scenario or as forensic evidence. This is to remind that this thesis is just a simulation of a forensic scenario where there is the need to discriminating writings in one-to-many communication. Therefore, this thesis may not be able to speak for typical FTC casework directly. However, it surely serves as a good gateway for future research endeavours that aim at introducing new scientific techniques based on the best-performing features reported in this thesis to remedy this situation. An attempt could also be made to form a database that are designed to be used in an FTC task, so research outputs based on the said database can say more about typical FTC cases.

9.4 Future Research

Although this thesis has come to an end, LR-based FTC is still far from an end since, there are a number of contributions that can be made to improve LR-based FTC.

This thesis has been experimenting an FTC task on product reviews which were written in all kinds of topics. The topic diversity of product reviews may more or less play a part in dictating which feature can perform robustly than the others. Moreover, when there is topic mismatch occurring in the documents of comparison as shown in Section 4.4, how a specific feature or feature type fares in this kind of situation is also of great interest in LR-based FTC.

This thesis has also been speculating regarding the optimal number of features for each feature set or experiment setting. There is even no speculation at all for the optimal numbers of PNGs for different word lengths. Research that can clearly illustrate how the FTC system performance changes as a function of the number of features would greatly confirm these speculations or answer some unanswered questions.

Another improvement that is immediately needed is the way to minimise the magnitude of counterfactual LRs while maintain, if not improve, that of the consistent-with-fact LRs for both pre-fusion and logistic regression fused LRs. As have been demonstrated through Tippett plots throughout this thesis, there often is the considerable magnitude of contrary-to-fact same-author and different-author LRs for every feature type and every word length. After the LRF, these errors could not still be fixed. The large magnitude of contrary-to-fact LRs only serves to contrast the consistent-with-fact LRs, which should be rectified at all cost so as to contribute to the improvements of LR-based FTC.

This thesis has also been talking about how the LRF makes the features interact within their types of features and how the types of features interact with one another in the fused FTC system between all the types of features. Nevertheless, it has not yet come to a satisfactory conclusion. Future research can show us more clearly how a feature interacts with another within a specific type of feature and how a feature type interacts with another when they all are logistic regression fused.

(21,994 Words Exclusive of Figures, Tables, References, and Appendices)

References

Aitken, C. G. G. (1995). Statistics and the Evaluation of Evidence for Forensic Scientists. Chichester: John Wiley.

Aitken, C. G. G., & Gold, E. (2013). Evidence evaluation for discrete data. *Forensic Science International, 230*(1-3), 147-155. doi:10.1016/j.forsciint.2013.02.042

Aitken, C. G. G., & Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society Series C-Applied Statistics, 53,* 109–122. http://dx.doi.org/10.1046/j.0035-9254.2003.05271.x.

Aitken, C. G. G., & Stoney, D. A. (1991). The Use of Statistics in Forensic Science. New York, London: Ellis Horwood.

Aitken, C. G. G., & Taroni, F. (2004). Statistics and the Evaluation of Evidence for Forensic Scientists. Chichester: Wiley.

Balding, D. J., & Steele, C. D. (2015). Weight-of-Evidence for Forensic DNA Profiles. Chichester: John Wiley & Sons. https://doi.org/10.1002/9781118814512

Bolck, A., Ni, H. F., & Lopatka, M. (2015). Evaluating score- and feature-based likelihood ratio models for multivariate continuous data: applied to forensic MDMA comparison. *Law, Probability and Risk, 14*(3), 243–266. https://doi.org/10.1093/lpr/mgv009

Bolck, A., Weyermann, C., Dujourdy, L. Esseiva, P., & van den Berg, J. (2009). Different likelihood ratio approaches to evaluate the strength of evidence of MDMA tablet comparisons. *Forensic Science International, 191*(1-3), 42-51. doi:10.1016/j.forsciint.2009.06.006

Belsivi, N. M. S., Muhammad. N., & Alonso-Fernandez, F. (2020, August 29-30). *Forensic Authorship Analysis of Microblogging Texts Using N-Grams and Stylometric Features*. [Paper presentation]. Proc. 8th International Workshop on Biometrics and Forensics (IWBF), Porto, Portugal. https://arxiv.org/abs/2003.11545

Benoit, K., & Matsuo, A. (2020). spacyr: Wrapper to the 'spaCy' 'NLP' Library (Version 1.2.1). The Comprehensive R Archive Network. Retrieved from https://cran.r-project.org/web/packages/spacyr/index.html

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An r package for the quantitative analysis of textual data (Version 2.1.2). The Comprehensive R Archive Network. Retrieved from https://quanteda.io

Brümmer, N., & du Preez, J. (2006). Application-independent evaluation of speaker detection. Computer Speech & Language, 20(2-3), 230-275. doi:10.1016/j.csl.2005.08.001

Carne, M., & Ishihara, S. (2020, Jan 14-15). *Feature-based forensic text comparison using a Poisson model for likelihood ratio estimation.* [Paper presentation]. The 18th Workshop of the Australasian Language Technology Association, Virtual Workshop. https://aclanthology.org/2020.alta-1.0

Chen, X. H., Champod, C., Yang, X., Shi, S. P., Luo, Y. W., Wang, N., & Lu, Q. M. (2018). Assessment of signature handwriting evidence via score-based likelihood ratio based on comparative measurement of relevant dynamic features. *Forensic Science International*, *282*, 101–110. https://doi.org/10.1016/j.forsciint.2017.11.022

Coyotl-Morales, R.M., Villaseñor-Pineda, L., Montes-y-Gómez, M., & Rosso, P. (2006). Authorship attribution using word sequences. iIn J.F. Martínez-Trinidad, J.A. Carrasco-Ochoa, & J. Kittler (Eds.), *Progress in Pattern Recognition, Image Analysis and Applications* (pp. 844-853).

Sanderson, C., & Guenter, S. (2006, Jul 22-23). *Short text authorship attribution via sequence kernels, Markov chains and author unmasking: an investigation*. [Paper presentation]. The 2006 Conference on Empirical Methods in Natural Language Processing, Sydney, NSW, Australia. https://dl.acm.org/doi/proceedings/10.5555/1610075 de Vel, O., Anderson A., Corney, M., & Mohay, G. (2001). Mining E-mail Content for Author Identification Forensics. *ACM SGM Record, 30*(4), 55-64. doi:10.1145/604264.604272

Diederich, J. (2003). Authorship Attribution with Support Vector Machines. *Applied Intelligence, 19,* 109-123. https://doi.org/10.1023/A:1023824908771

Eloy, J. A., Li, S., Kasabwala, K., Agarwal, N., Hansberry, D. R., Baredes, S., & Setzen, M. (2012). Readability Assessment of Patient Education Materials on Major Otolaryngology Association Websites. *Otolaryngology–Head and Neck Surgery, 147*(5), 848-854. https://doi-org.virtual.anu.edu.au/10.1177/0194599812456152

Evert, S., Proisl, T., Jannidis, F., Pielström, S., Schöch, C., & Vitt, T. (2015, June 4). *Towards a better understanding of Burrows's Delta in literary authorship attribution.* [Paper presentation]. NAACL-HLT Fourth Workshop on Computational Linguistic for Literature, Denver, Colorado. doi:10.3115/v1/W15-0709

Evert, S., Proisl, T., Jannidis, F., Reger, I., Pielström, S., Schöch, C., & Vitt, T. (2017). Understanding and explaining Delta measures for authorship attribution. *Digital Scholarship in Humanities, 32*(2), ii4-ii16. doi:10.1093/llc/fqx023

Evett, I. W. (1998). Towards a uniform framework for reporting opinions in forensic science casework. *Science & Justice, 38*(3), 198–202. https://dx.doi.org/10.1016/S1355-0306(98)72105-7

Forsyth, R. S., & Holmes, D. I. (1996). Feature-finding for text classification. *Literary and Linguistic Computing, 11*(4), 163–174. doi:http://dx.doi.org/10.1093/llc/11.4.163.

Flesch, R. (1948). A new readability yardstick *Journal of Applied Psychology, 32*(3), 221-233. doi:10.1037/h0057532

Gamon, M. (2004, Aug 23-27). *Authorship classification with deep linguistic analysis features.* [Paper presentation]. The 20th International Conference on Computational Linguistics, Geneva, Switzerland. https://aclanthology.org/C04-1088

Garton, N., Ommen, D., Niemi, J., & Carriquiry, A. (2020). Score-based likelihood ratios to evaluate forensic pattern evidence. Retrieved from https://arxiv.org/abs/2002.09470.

Grant, T. (2007) Quantifying evidence in forensic authorship analysis. *International Journal of Speech Language and the Law, 14*(1), 1–25. http://dx.doi.org/10.1558/ijsll.v14i1.1.

Grant, T. (2010) Text messaging forensics: txt 4n6: Idiolect free authorship analysis? In A. J. M. Coulthard (Eds.), *The Routledge Handbook of Forensic Linguistics* (pp. 508–522). Abingdon: Routledge.

Grieve, J. (2007). Quantitative authorship attribution: an evaluation of techniques. *Literary and Linguistic Computing, 22*(3), 251–270. doi:http://dx.doi.org/10.1093/llc/fqm020.

Halvani, O., Winter, C., & Graner, L. (2017). AV Corpora [Data file]. Retreived from https://www.dropbox.com/sh/qelav1ap0uwo82q/AADcyI77qrv-ctJosCOhwMRUa?dl=0

Halvani, O., Winter, C., & Graner, L. (2017). Authorship verification based on compression-models. *arXiv preprint arXiv:1706.00516.* Retrieved on 5 May 2021 from http://arxiv.org/abs/1706.00516.

He, R., & McAuley, J. (2016, Apr 11-15). *Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering*. [Paper presentation]. World Wide Web Conference, Montreal, Quebec, Canada. doi:10.1145/2872427.2883037

Helper, A. B., Saunders, C. P., Davis, L. J., & Buscaglia, J. (2012) Score-based likelihood ratios for handwriting evidence. *Forensic Science International, 219*(3), 129-140. doi:10.1016/j.forsciint.2011.12.009

Hicks, T., Biedermann, A., de Koeijer, J. A., Taroni, F., Champod, C., & Evett, I. W. (2015). The importance of distinguishing information from evidence/observations when formulating propositions. *Science & Justice, 55*(6), 520-525. doi:10.1016/j.scijus.2015.06.008

Houvardas, J., & Stamatatos, E. (2006, Sep 12-15). *N-gram Feature Selection for Authorship Identification.* [Paper presentation]. Artificial Intelligence: Methodology, Systems, and Applications, 12th International Conference (AIMSA 2006), Varna, Bulgaria. https://link.springer.com/chapter/10.1007/11861461_10

Iqbal, F., Khan, L. A., Fung, B. C. M., & Debbadi, M. (2010, Mar 22-26). *E-mail Authorship Verification for Forensic Investigation*. [Paper presentation]. The 2010 ACM Symposium on Applied Computing, Sierre, Switzerland. https://dl-acm-org.virtual.anu.edu.au/doi/10.1145/1774088.1774428

Ishihara, S. (2014). A likelihood ratio-based evaluation of strength of authorship attribution evidence in SMS messages using N-grams. *International Journal of Speech Language and the Law, 21*(1). doi:10.1558/ijsll.v21i1.23

Ishihara, S. (2017a). Strength of forensic text comparison evidence from stylometric features: a multivariate likelihood ratio-based analysis. *International Journal of Speech Language and the Law, 24*(1), 67-98. doi:10.1558/ijsll.30305

Ishihara, S. (2017b). Strength of linguistic text evidence: A fused forensic text comparison system. *Forensic Science International, 278*, 184-197. doi:10.1016/j.forsciint.2017.06.040

Ishihara, S. (2021). Score-based likelihood ratios for linguistic text evidence with a bag-of-words model. *Forensic Science International, 327*, 110980. doi:10.1016/j.forsciint.2021.110980

Johnston, J., & Giles, M. (2017). Still Flumoxed by the Flesch Kincaid Score?: Readability and Plain English in Surgical Patient Information Leaflets. *International Journal of Surgery, 47,* S55. https://doi.org/10.1016/j.ijsu.2017.08.285

Keselj, V., Peng, K., Cercone, N., & Thomas, C. (2003, Aug 22-25). *N-GRAM-BASED AUTHOR PROFILES FOR AUTHORSHIP ATTRIBUTION*. [Paper presentation]. The Conference Pacific Association for Computational Linguistics (PACLING'03), Halifax, Nova Scotia, Canada. https://web.cs.dal.ca/~vlado/papers/meta/Kes03.html

Kincald, J. P., Fishburne Jr., R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*. Retrieved from https://stars.library.ucf.edu/istlibrary/56/

Koppel, M., Schler, J., & Argamon, S. (2011). Authorship Attribution in the wild. *Language Resources and Education, 45,* 83-94. doi:10.1007/s10579-009-9111-2

Kukushkina, O. V., Polikarpov, A. A., & Khmelev, D. V. (2001). Using Literal and Grammatical Statistics for Authorship Attribution. *Problems of Information Transmission, 37*, 172-184. https://doi.org/10.1023/A:1010478226705

Lambers, M., & Veenman, C. J. (2009). Forensic authorship attribution using compression distances to prototypes. In Z. Geradts, K. Y. Franke, & C. J. Veenman (Eds.), *Computational Forensics* (pp. 13–24). Berlin: Springer.

Leegwater, A. J., Meuwly, D., Sjerps, M., Vergeer, P., & Alberink, I. (2017). Performance study of a score-based likelihood ratio system for forensic fingermark comparison. *Journal of Forensic Sciences*. *62*(3), 626–640. https://doi.org/10.1111/1556- 4029.13339

Lund, S. P., & Iyer, H. (2017). Likelihood Ratio as Weight of Forensic Evidence: A Closer Look. *Journal of Research of National Institute of Standards and Technology, 122*(27), 1-32. https://doi.org/10.6028/jres.122.027

McMenamin, G. R. (2001). Style markers in authorship studies. *International Journal of Speech Languange and the Law*, *8*(2), 93–97. https://doi.org/10.1558/sll.2001.8.2.93

McMenamin, G. R. (2001). Forensic Linguistics: Advances in Forensic Stylistics. Boca Raton: CRC Press

Morrison, G. S. (2013). Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences, 45*(2), 173-197. https://dx.doi.org/10.1080/00450618.2012.733025

Morrison, G. S., & Enzinger, E. (2018). Score based procedures for the calculation of forensic likelihood ratios - scores should take account of both similarity and typicality. *Science & Justice*, *58*(1), 47–58. https://doi.org/10.1016/j.scijus.2017.06.005

Mutinda, J., Mwangi, W., & George, O. (2021). Lexicon-pointed hybrid N-gram Features Extraction Model (LeNFEM) for sentence level sentiment analysis. *Engineering Reports, 3*(8), 1-17. https://doi.org/10.1002/eng2.12374

Neumann, C., & Ausdemore, M. (2020). Defence against the modern arts: the curse of statistics-Part II: 'Score-based likelihood ratios'. *Law, Probability and Risk, 19*(1), 21–42. https://doi.org/10.1093/lpr/mgaa006

Ooms, J. (2020). hunspell: High-performance stemmer, tokenizer, and spell checker (Version 3.0.1). The Comprehensive R Archive Network. Retrieved from https://CRAN.R-project.org/package=hunspell

Peng, F., Schuurmans, D., Wang, S., & Keselj, V. (2003, Apr 12-17). *Language Independent Authorship Attribution using Character Level Language Models*. [Paper presentation]. The 10th Conference of the European Chapter of the Association for Computational Linguistics, Budapest, Hungary. https://aclanthology.org/E03-1053/

Robertson, B., & Vignaux, G. A. (1995). Interpreting Evidence: Evaluating Forensic Science in the Courtroom. Chichester: John Wiley & Sons.

Rocha, A., Scheirer, W. J., Forstall, C. W., Cavalcante, T., Theophilo, A., Shen, B., . . . Stamatatos, E. (2017). Authorship Attribution for Social Media Forensics. *IEEE Transactions on Information Forensics and Security, 12*(1), 5-33. doi:10.1109/tifs.2016.2603960

Rudman, J. (1997). The state of authorship attribution studies: some problems and solutions. *Computers and the Humanities*, 31(4), 351–365. doi:http://dx.doi.org/10.1023/A:1001018624850.

Saks, M. J., & Koehler, J. J. (2005). The coming paradigm shift in forensic identification science. Science, 309(5736), 892-895. doi:10.1126/science.1111565

Sari, Y., Stevenson, M., & Vlachos, A. (2018, Aug 20-26). *Topic or Style? Exploring the Most Useful Features for Authorship Attribution.* [Paper presentation]. The 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA. https://aclanthology.org/C18-1029/

Savoy, J. (2012a). Authorship Attribution: A Comparative Study of Three Text Corpora and Three Languages. *Journal of Quantitative Linguistics, 19*(2), 132-161. https://doi.org/10.1080/09296174.2012.659003

Savoy, J. (2012b). Authorship Attribution Based on Specific Vocabulary. *ACM Transactions on Information Systems, 30*(2), 1-30. https://doi.org/10.1145/2180868.2180874

Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., & Chanona-Hernandez, L. (2014). Syntactic N-grams as machine learning features for natural language processing. *Expert Systems with Applications: An International Journal. 41*(3), 853-860. https://doi.org/10.1016/j.eswa.2013.08.015

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology, 60*(3), 538-556. doi:10.1002/asi.21001

Stamatatos, E. (2013). On the Robustness of Authorship Attribution Based on Character N-gram features. *Journal of Law and Policy, 21*(2), 421-440. Retrieved from https://www.semanticscholar.org/paper/On-the-Robustness-of-Authorship-Attribution-Based-N-Stamatatos/b3150811a2a38afb00ed264aa8c039d05a7199ce

Stefanovič, P., Olga, K., & Štrimaitis. R. (2019). The N-Grams Based Text Similarity Detection Approach Using Self-Organizing Maps and Similarity Measures. *Applied Sciences, 9*(9), 1-15. doi:10.3390/app9091870

Tambouli, M., & Prasad, R. (2019). A robust authorship attribution on big period. nternational *Journal of Electrical and Computer Engineering, 9*(4), 3167-3174. doi: 10.11591/ijece.v9i4.pp3167-3174

Tweedia, J. F., & Baayen. R. H. (1998). How Variable May a Constant Be Measures of Lexical Richness in Perspective. *Computers and the Humanities, 32*(5), 323-352. doi:10.1023/A:1001749303137

Vergeer, P., van Es, A., de Jongh, A., Alberink, I., & Stoel, R. (2016). Numerical likelihood ratios outputted by LR systems are often based on extrapolation: When to stop extrapolating? Science & Justice, 56(6), 482-491. doi:10.1016/j.scijus.2016.06.003

Yule, G. U. (1944). *The Statistical Study of Literary Vocabulary*. Cambridge, UK: Cambridge University Press.

Zheng, R., Li, J., Chen, H., & Huang, Z. (2006). A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology, 57*(3), 378-393. doi:10.1002/asi.2031

Zheng, R., Qin, Y., Huang, Z., & Chen, H. C. (2003, Jun 2-3). *Authorship analysis in cybercrime investigation.* [Paper presentation]. The 1st NSF/NIJ Conference on Intelligence and Security Informatics, Tuzcon, AZ, USA. https://dl.acm.org/doi/proceedings/10.5555/1792094

Appendix A Statistical Features: Full Results

Table A1

Best-performing Combinations of Two Statistical Features for 700, 1,400, and 2,100 Word Lengths

Sample Word Number	Pair of Features	*Cllr*		$Cllr^{min}$		$Cllr^{cal}$		*EER*	
		Pre	Post	Pre	Post	Pre	Post	Pre	Post
700	(4,6)	0.909	0.907	0.887	0.887	0.022	0.021	0.330	0.330
	(6,9)	0.917	0.918	0.893	0.893	0.024	0.025	0.328	0.328
	(4,7)	0.934	0.931	0.910	0.910	0.024	0.021	0.346	0.346
	(4,9)	0.937	0.939	0.910	0.910	0.027	0.028	0.355	0.355
	(7,9)	0.938	0.937	0.906	0.926	0.032	0.019	0.906	0.367
	(1,6)	0.945	0.940	0.916	0.916	0.029	0.024	0.350	0.350
	(3,6)	0.945	0.942	0.921	0.921	0.024	0.020	0.358	0.358
	(1,4)	0.948	0.945	0.926	0.926	0.022	0.019	0.367	0.367
	(1,7)	0.953	0.950	0.928	0.928	0.025	0.022	0.369	0.369
	(1,9)	0.955	0.954	0.936	0.936	0.019	0.017	0.374	0.374
1,400	(6,7)	0.839	0.838	0.811	0.811	0.028	0.027	0.276	0.276
	(4,6)	0.840	0.833	0.807	0.807	0.033	0.025	0.287	0.287
	(6,9)	0.858	0.858	0.824	0.824	0.035	0.034	0.293	0.293
	(1,6)	0.867	0.867	0.834	0.834	0.033	0.033	0.296	0.296
	(4,7)	0.876	0.868	0.848	0.848	0.027	0.019	0.298	0.298
	(7,9)	0.877	0.875	0.847	0.847	0.030	0.028	0.300	0.300
	(1,7)	0.887	0.888	0.858	0.858	0.029	0.030	0.316	0.316
	(3,6)	0.888	0.887	0.859	0.859	0.029	0.028	0.307	0.307
	(1,4)	0.891	0.891	0.869	0.869	0.022	0.022	0.316	0.316
	(4,9)	0.893	0.893	0.858	0.858	0.035	0.035	0.319	0.319
2,100	(6,7)	0.777	0.776	0.741	0.741	0.036	0.036	0.247	0.247
	(4,6)	0.792	0.787	0.764	0.764	0.027	0.022	0.255	0.255
	(6,9)	0.801	0.773	0.028	0.261	0.801	0.773	0.027	0.261
	(1,6)	0.809	0.809	0.783	0.783	0.026	0.026	0.264	0.264
	(4,7)	0.826	0.821	0.802	0.802	0.024	0.019	0.279	0.279
	(7,9)	0.829	0.826	0.796	0.796	0.033	0.030	0.264	0.264
	(3,6)	0.831	0.829	0.803	0.803	0.028	0.026	0.274	0.274
	(1,7)	0.840	0.839	0.818	0.818	0.022	0.021	0.286	0.286
	(3,7)	0.852	0.852	0.825	0.825	0.028	0.027	0.288	0.288
	(1,4)	0.856	0.856	0.834	0.834	0.021	0.022	0.285	0.285

Table A2

Best-performing Combinations of Three Statistical Features for 700, 1,400, and 2,100 Word Lengths

Sample Word Number	Combination of features	Cllr	$Cllr^{min}$	$Cllr^{cal}$	EER
700	(6,7,9)	0.862	0.831	0.032	0.293
	(4,6,7)	0.872	0.837	0.035	0.300
	(4,6,9)	0.875	0.849	0.026	0.304
	(1,6,7)	0.889	0.859	0.030	0.317
	(6,7,8)	0.889	0.857	0.032	0.313
	(3,6,7)	0.891	0.859	0.032	0.315
	(4,6,8)	0.896	0.867	0.029	0.316
	(6,8,9)	0.896	0.869	0.028	0.322
	(1,4,6)	0.901	0.865	0.035	0.322
	(4,7,9)	0.901	0.868	0.033	0.323
1,400	(6,7,9)	0.762	0.729	0.034	0.244
	(4,6,7)	0.763	0.729	0.035	0.242
	(1,6,7)	0.777	0.741	0.036	0.254
	(4,6,9)	0.785	0.746	0.039	0.255
	(1,4,6)	0.790	0.756	0.034	0.257
	(3,6,7)	0.794	0.762	0.032	0.260
	(6,7,8)	0.797	0.765	0.031	0.263
	(4,6,8)	0.808	0.776	0.032	0.270
	(1,6,8)	0.810	0.779	0.030	0.265
	(6,8,9)	0.812	0.775	0.037	0.277
2,100	(6,7,9)	0.680	0.652	0.028	0.203
	(4,6,7)	0.689	0.659	0.030	0.214
	(1,6,7)	0.694	0.667	0.027	0.215
	(3,6,7)	0.713	0.685	0.028	0.222
	(4,6,9)	0.720	0.685	0.035	0.226
	(1,4,6)	0.728	0.703	0.025	0.229
	(6,7,8)	0.730	0.696	0.034	0.232
	(3,4,6)	0.742	0.715	0.027	0.233
	(6,8,9)	0.745	0.716	0.029	0.236
	(1,6,8)	0.745	0.720	0.026	0.239

Table A3

Best-performing Combinations of Four Statistical Features for 700, 1,400, and 2,100 Word Lengths

Sample Word Number	Combination of features	*Cllr*	*Cllr^{min}*	*Cllr^{cal}*	*EER*
700	(4,6,7,9)	0.830	0.802	0.027	0.283
	(6,7,8,9)	0.838	0.813	0.025	0.293
	(4,6,7,8)	0.847	0.823	0.024	0.292
	(1,4,6,7)	0.849	0.819	0.030	0.299
	(3,6,7,9)	0.852	0.829	0.023	0.300
	(1,6,7,9)	0.852	0.826	0.026	0.300
	(3,4,6,7)	0.854	0.825	0.029	0.298
	(4,6,8,9)	0.859	0.834	0.026	0.300
	(6,7,9,10)	0.860	0.841	0.019	0.305
	(5,6,7,9)	0.861	0.836	0.024	0.298
1,400	(4,6,7,9)	0.711	0.683	0.029	0.222
	(1,4,6,7)	0.715	0.689	0.027	0.233
	(6,7,8,9)	0.724	0.696	0.028	0.235
	(1,6,7,9)	0.730	0.698	0.032	0.241
	(4,6,7,8)	0.733	0.707	0.026	0.239
	(1,6,7,8)	0.733	0.708	0.025	0.237
	(3,4,6,7)	0.735	0.711	0.024	0.236
	(3,6,7,9)	0.739	0.711	0.028	0.243
	(5,6,7,9)	0.742	0.721	0.021	0.239
	(2,6,7,9)	0.743	0.724	0.020	0.244
2,100	(4,6,7,9)	0.626	0.602	0.024	0.191
	(6,7,8,9)	0.639	0.612	0.027	0.190
	(1,4,6,7)	0.640	0.616	0.024	0.196
	(1,6,7,9)	0.649	0.628	0.022	0.199
	(1,6,7,8)	0.654	0.631	0.023	0.201
	(3,4,6,7)	0.655	0.632	0.023	0.202
	(3,6,7,9)	0.662	0.642	0.020	0.203
	(4,6,7,8)	0.662	0.631	0.031	0.207
	(5,6,7,9)	0.665	0.645	0.020	0.211
	(3,6,7,8)	0.669	0.644	0.025	0.205

Table A4

Best-performing Combinations of Five Statistical Features for 700, 1,400, and 2,100 Word Lengths

Sample Word Number	Combination of features	*Cllr*	*Cllr^{min}*	*Cllr^{cal}*	*EER*
700	(4,6,7,8,9)	0.814	0.791	0.022	0.283
	(1,4,6,7,9)	0.826	0.800	0.026	0.287
	(3,4,6,7,9)	0.826	0.804	0.022	0.286
	(4,5,6,7,9)	0.834	0.811	0.024	0.286
	(4,6,7,9,10)	0.834	0.817	0.017	0.292
	(3,6,7,8,9)	0.835	0.816	0.019	0.290
	(1,6,7,8,9)	0.835	0.814	0.022	0.295
	(1,4,6,7,8)	0.835	0.809	0.026	0.294
	(3,4,6,7,8)	0.837	0.815	0.022	0.290
	(2,4,6,7,9)	0.839	0.820	0.019	0.298
1,400	(4,6,7,8,9)	0.688	0.663	0.025	0.216
	(1,4,6,7,8)	0.689	0.665	0.025	0.221
	(1,4,6,7,9)	0.692	0.666	0.026	0.225
	(1,6,7,8,9)	0.701	0.675	0.026	0.226
	(3,4,6,7,9)	0.703	0.677	0.026	0.225
	(4,5,6,7,9)	0.704	0.685	0.019	0.224
	(2,4,6,7,9)	0.705	0.686	0.019	0.227
	(3,4,6,7,8)	0.707	0.681	0.025	0.218
	(3,6,7,8,9)	0.710	0.686	0.024	0.225
	(1,4,5,6,7)	0.716	0.695	0.021	0.233
2,100	(4,6,7,8,9)	0.601	0.578	0.023	0.180
	(1,4,6,7,9)	0.612	0.592	0.020	0.187
	(1,4,6,7,8)	0.616	0.591	0.025	0.190
	(1,6,7,8,9)	0.621	0.597	0.024	0.187
	(3,4,6,7,9)	0.625	0.606	0.019	0.192
	(4,5,6,7,9)	0.626	0.606	0.020	0.199
	(3,4,6,7,8)	0.631	0.606	0.024	0.190
	(2,4,6,7,9)	0.631	0.610	0.021	0.202
	(3,6,7,8,9)	0.633	0.607	0.026	0.188
	(2,6,7,8,9)	0.636	0.618	0.018	0.199

Table A5

Best-performing Combinations of Six Statistical Features for 700, 1,400, and 2,100 Word Lengths

Sample Word Number	Combination of features	*Cllr*	*Cllr^{min}*	*Cllr^{cal}*	*EER*
700	(3,4,6,7,8,9)	0.814	0.792	0.022	0.277
	(1,4,6,7,8,9)	0.816	0.791	0.025	0.282
	(4,5,6,7,8,9)	0.822	0.800	0.022	0.276
	(2,4,6,7,8,9)	0.825	0.803	0.022	0.295
	(4,6,7,8,9,10)	0.826	0.809	0.017	0.288
	(1,4,5,6,7,9)	0.827	0.806	0.022	0.289
	(1,4,6,7,9,10)	0.828	0.812	0.016	0.290
	(3,4,5,6,7,9)	0.829	0.808	0.020	0.284
	(3,4,6,7,9,10)	0.831	0.815	0.016	0.292
	(1,2,4,6,7,9)	0.832	0.813	0.019	0.297
1,400	(1,4,6,7,8,9)	0.671	0.650	0.022	0.211
	(3,4,6,7,8,9)	0.682	0.657	0.024	0.211
	(2,4,6,7,8,9)	0.683	0.666	0.017	0.216
	(1,4,5,6,7,9)	0.688	0.671	0.017	0.217
	(1,2,4,6,7,9)	0.688	0.668	0.021	0.227
	(4,5,6,7,8,9)	0.689	0.672	0.016	0.214
	(1,2,4,6,7,8)	0.691	0.672	0.020	0.228
	(2,3,4,6,7,9)	0.694	0.673	0.021	0.229
	(3,4,5,6,7,9)	0.696	0.678	0.018	0.222
	(1,4,5,6,7,8)	0.696	0.678	0.018	0.220
2,100	(1,4,6,7,8,9)	0.594	0.571	0.023	0.180
	(2,4,6,7,8,9)	0.605	0.586	0.019	0.185
	(3,4,6,7,8,9)	0.606	0.581	0.025	0.181
	(4,5,6,7,8,9)	0.610	0.592	0.017	0.192
	(1,4,5,6,7,9)	0.610	0.591	0.019	0.196
	(1,2,4,6,7,9)	0.617	0.594	0.022	0.200
	(3,4,5,6,7,9)	0.618	0.601	0.017	0.199
	(1,5,6,7,8,9)	0.619	0.600	0.019	0.198
	(1,4,6,7,9,10)	0.620	0.603	0.018	0.191
	(1,2,6,7,8,9)	0.621	0.600	0.020	0.195

Table A6

Best-performing Combinations of Seven Statistical Features for 700, 1,400, and 2,100 Word Lengths

Sample Word Number	Combination of features	*Cllr*	*Cllr^{min}*	*Cllr^{cal}*	*EER*
700	(1,4,5,6,7,8,9)	0.817	0.797	0.020	0.281
	(3,4,5,6,7,8,9)	0.818	0.799	0.019	0.278
	(1,2,4,6,7,8,9)	0.820	0.800	0.020	0.289
	(2,3,4,6,7,8,9)	0.821	0.804	0.017	0.285
	(1,4,6,7,8,9,10)	0.822	0.805	0.017	0.288
	(2,4,5,6,7,8,9)	0.824	0.804	0.021	0.293
	(3,4,6,7,8,9,10)	0.824	0.808	0.016	0.289
	(1,4,5,6,7,9,10)	0.829	0.813	0.016	0.291
	(1,2,4,5,6,7,9)	0.830	0.812	0.018	0.296
	(2,3,4,5,6,7,9)	0.830	0.812	0.018	0.288
1,400	(1,2,4,6,7,8,9)	0.670	0.648	0.021	0.218
	(1,4,5,6,7,8,9)	0.671	0.655	0.016	0.210
	(2,3,4,6,7,8,9)	0.674	0.656	0.018	0.217
	(3,4,5,6,7,8,9)	0.680	0.662	0.018	0.216
	(1,2,4,5,6,7,9)	0.683	0.667	0.016	0.223
	(2,4,5,6,7,8,9)	0.684	0.669	0.015	0.220
	(2,3,4,5,6,7,9)	0.688	0.668	0.020	0.226
	(1,4,6,7,8,9,10)	0.689	0.667	0.022	0.225
	(1,2,5,6,7,8,9)	0.694	0.677	0.017	0.229
	(1,2,4,5,6,7,8)	0.695	0.676	0.019	0.230
2,100	(1,4,5,6,7,8,9)	0.595	0.575	0.020	0.188
	(1,2,4,6,7,8,9)	0.596	0.575	0.021	0.186
	(2,3,4,6,7,8,9)	0.601	0.579	0.022	0.182
	(3,4,5,6,7,8,9)	0.603	0.582	0.021	0.192
	(1,4,6,7,8,9,10)	0.606	0.588	0.019	0.189
	(2,4,5,6,7,8,9)	0.610	0.593	0.017	0.189
	(1,2,4,5,6,7,9)	0.612	0.590	0.022	0.198
	(2,4,6,7,8,9,10)	0.614	0.599	0.014	0.192
	(3,4,6,7,8,9,10)	0.614	0.594	0.020	0.189
	(1,2,4,6,7,9,10)	0.615	0.596	0.019	0.191

Table A7

Best-performing Combinations of Eight Statistical Features for 700, 1,400, and 2,100 Word Lengths

Sample Word Number	Combination of features	*Cllr*	*Cllr^{min}*	*Cllr^{cal}*	*EER*
700	(2,3,4,5,6,7,8,9)	0.819	0.801	0.018	0.285
	(1,2,4,5,6,7,8,9)	0.819	0.800	0.020	0.292
	(1,4,5,6,7,8,9,10)	0.823	0.808	0.015	0.287
	(3,4,5,6,7,8,9,10)	0.826	0.810	0.016	0.287
	(1,2,4,6,7,8,9,10)	0.828	0.811	0.017	0.294
	(2,3,4,6,7,8,9,10)	0.829	0.809	0.020	0.293
	(2,4,5,6,7,8,9,10)	0.829	0.812	0.018	0.298
	(1,2,4,5,6,7,9,10)	0.831	0.816	0.014	0.295
	(2,3,4,5,6,7,9,10)	0.831	0.816	0.015	0.292
	(1,3,4,5,6,7,8,9)	0.832	0.812	0.021	0.288
1,400	(1,2,4,5,6,7,8,9)	0.669	0.652	0.017	0.217
	(2,3,4,5,6,7,8,9)	0.673	0.653	0.020	0.222
	(1,2,4,6,7,8,9,10)	0.681	0.663	0.018	0.224
	(2,3,4,6,7,8,9,10)	0.685	0.667	0.018	0.222
	(1,4,5,6,7,8,9,10)	0.686	0.672	0.014	0.219
	(1,2,3,4,6,7,8,9)	0.689	0.670	0.019	0.221
	(1,2,4,5,6,7,9,10)	0.690	0.673	0.016	0.224
	(3,4,5,6,7,8,9,10)	0.691	0.676	0.015	0.219
	(2,3,4,5,6,7,9,10)	0.692	0.678	0.014	0.226
	(1,3,4,5,6,7,8,9)	0.693	0.672	0.021	0.222
2,100	(1,2,4,5,6,7,8,9)	0.597	0.575	0.022	0.188
	(2,3,4,5,6,7,8,9)	0.601	0.579	0.022	0.189
	(1,2,4,6,7,8,9,10)	0.602	0.584	0.018	0.186
	(2,3,4,6,7,8,9,10)	0.605	0.586	0.019	0.185
	(1,4,5,6,7,8,9,10)	0.606	0.589	0.017	0.189
	(3,4,5,6,7,8,9,10)	0.612	0.596	0.016	0.191
	(1,2,4,5,6,7,9,10)	0.613	0.591	0.022	0.190
	(2,3,4,5,6,7,9,10)	0.615	0.594	0.021	0.191
	(2,4,5,6,7,8,9,10)	0.616	0.601	0.016	0.191
	(1,2,3,4,6,7,8,9)	0.616	0.599	0.018	0.190

Table A8

Best-performing Combinations of Nine Statistical Features for 700, 1,400, and 2,100 Word Lengths

Sample Word Number	Combination of features	*Cllr*	*Cllr^{min}*	*Cllr^{cal}*	*EER*
700	(1,2,4,5,6,7,8,9,10)	0.825	0.809	0.016	0.293
	(2,3,4,5,6,7,8,9,10)	0.825	0.808	0.017	0.291
	(1,2,3,4,5,6,7,8,9)	0.830	0.811	0.019	0.292
	(1,3,4,5,6,7,8,9,10)	0.836	0.818	0.018	0.295
	(1,2,3,4,6,7,8,9,10)	0.838	0.821	0.017	0.300
	(1,2,3,4,5,6,7,9,10)	0.840	0.824	0.016	0.299
	(1,2,3,5,6,7,8,9,10)	0.850	0.835	0.015	0.309
	(1,2,3,4,5,6,7,8,10)	0.850	0.834	0.016	0.306
	(1,2,3,4,5,6,8,9,10)	0.863	0.846	0.017	0.315
	(1,2,3,4,5,7,8,9,10)	0.868	0.852	0.015	0.317
1,400	(1,2,4,5,6,7,8,9,10)	0.678	0.664	0.014	0.218
	(2,3,4,5,6,7,8,9,10)	0.681	0.665	0.015	0.219
	(1,2,3,4,5,6,7,8,9)	0.684	0.664	0.020	0.224
	(1,2,3,4,6,7,8,9,10)	0.695	0.676	0.019	0.227
	(1,3,4,5,6,7,8,9,10)	0.699	0.682	0.017	0.229
	(1,2,3,4,5,6,7,9,10)	0.700	0.682	0.018	0.231
	(1,2,3,5,6,7,8,9,10)	0.708	0.691	0.018	0.232
	(1,2,3,4,5,6,7,8,10)	0.712	0.693	0.019	0.234
	(1,2,3,4,5,6,8,9,10)	0.727	0.710	0.017	0.246
	(1,2,3,4,5,7,8,9,10)	0.742	0.727	0.016	0.247
2,100	(1,2,4,5,6,7,8,9,10)	0.601	0.581	0.020	0.183
	(2,3,4,5,6,7,8,9,10)	0.604	0.585	0.018	0.185
	(1,2,3,4,5,6,7,8,9)	0.613	0.591	0.022	0.193
	(1,2,3,4,6,7,8,9,10)	0.615	0.599	0.016	0.188
	(1,3,4,5,6,7,8,9,10)	0.620	0.603	0.017	0.195
	(1,2,3,4,5,6,7,9,10)	0.623	0.602	0.021	0.193
	(1,2,3,5,6,7,8,9,10)	0.629	0.611	0.017	0.197
	(1,2,3,4,5,6,7,8,10)	0.636	0.618	0.018	0.195
	(1,2,3,4,5,6,8,9,10)	0.657	0.639	0.018	0.210
	(1,2,3,4,5,7,8,9,10)	0.678	0.659	0.018	0.221

Appendix B Word N-gram Features: Full Results

Table B1

Best-performing Experiment Settings for Word Unigrams (n=1) for 700, 1,400, and 2,100 Word Lengths

Sample Word Number	Number of Features	*Cllr*	*Cllr*^{min}	*Cllr*^{cal}	*EER*
700	300	0.711	0.695	0.016	0.235
	275	0.712	0.697	0.015	0.241
	325	0.721	0.704	0.016	0.235
	350	0.728	0.711	0.017	0.242
	225	0.728	0.712	0.016	0.239
	250	0.728	0.716	0.013	0.244
	125	0.728	0.714	0.015	0.251
	450	0.729	0.712	0.017	0.235
	375	0.729	0.711	0.018	0.240
	175	0.729	0.715	0.014	0.241
1,400	300	0.438	0.424	0.014	0.129
	275	0.441	0.427	0.014	0.127
	325	0.449	0.435	0.014	0.131
	350	0.452	0.441	0.011	0.133
	375	0.457	0.446	0.011	0.136
	250	0.457	0.443	0.014	0.132
	400	0.46	0.445	0.015	0.136
	225	0.461	0.448	0.013	0.137
	450	0.462	0.45	0.013	0.142
	425	0.466	0.452	0.014	0.142
2,100	300	0.302	0.292	0.010	0.086
	275	0.305	0.294	0.011	0.082
	350	0.308	0.297	0.010	0.083
	325	0.308	0.298	0.010	0.083
	375	0.311	0.300	0.011	0.084
	400	0.318	0.307	0.011	0.085
	450	0.320	0.308	0.012	0.087
	225	0.321	0.308	0.012	0.089
	250	0.322	0.308	0.014	0.087
	425	0.323	0.311	0.012	0.091

Table B2

Best-performing Experiment Settings for Word Bigrams (n=2) for 700, 1,400, and 2,100 Word Lengths

Sample Word Number	Number of Features	*Cllr*	*Cllr^{min}*	*Cllr^{cal}*	*EER*
700	575	0.763	0.738	0.025	0.251
	550	0.765	0.741	0.024	0.253
	600	0.766	0.741	0.025	0.254
	525	0.767	0.745	0.022	0.255
	500	0.772	0.749	0.023	0.262
	450	0.775	0.751	0.023	0.257
	475	0.776	0.750	0.026	0.260
	200	0.776	0.759	0.017	0.275
	425	0.777	0.753	0.024	0.262
	175	0.778	0.758	0.019	0.268
1,400	575	0.464	0.453	0.011	0.139
	600	0.464	0.452	0.012	0.140
	550	0.467	0.455	0.012	0.140
	525	0.468	0.453	0.015	0.142
	475	0.477	0.465	0.013	0.146
	500	0.478	0.463	0.014	0.147
	450	0.480	0.468	0.011	0.145
	375	0.485	0.473	0.012	0.144
	425	0.486	0.470	0.016	0.146
	400	0.487	0.474	0.012	0.147
2,100	600	0.312	0.300	0.012	0.083
	575	0.314	0.300	0.014	0.083
	550	0.314	0.301	0.014	0.085
	525	0.319	0.305	0.014	0.087
	500	0.323	0.311	0.012	0.088
	475	0.328	0.316	0.011	0.094
	450	0.329	0.317	0.011	0.094
	425	0.334	0.322	0.012	0.090
	400	0.342	0.332	0.010	0.096
	375	0.344	0.332	0.011	0.097

Table B3

Best-performing Experiment Settings for Word Trigrams (n=3) for 700, 1,400, and 2,100 Word Lengths

Sample Word Number	Number of Features	*Cllr*	*Cllr^{min}*	*Cllr^{cal}*	*EER*
700	425	0.875	0.863	0.013	0.332
	600	0.876	0.862	0.014	0.327
	450	0.877	0.864	0.013	0.333
	475	0.877	0.863	0.014	0.331
	550	0.878	0.861	0.017	0.329
	500	0.878	0.864	0.015	0.329
	575	0.879	0.863	0.015	0.331
	525	0.879	0.865	0.015	0.330
	350	0.879	0.867	0.012	0.330
	400	0.880	0.867	0.013	0.335
1,400	550	0.668	0.653	0.015	0.221
	575	0.671	0.655	0.016	0.220
	525	0.674	0.660	0.013	0.224
	600	0.675	0.661	0.013	0.221
	500	0.675	0.663	0.012	0.223
	475	0.675	0.662	0.013	0.220
	450	0.678	0.663	0.015	0.224
	425	0.680	0.667	0.013	0.227
	400	0.690	0.677	0.013	0.230
	350	0.692	0.679	0.013	0.230
2,100	550	0.526	0.507	0.019	0.153
	575	0.528	0.512	0.016	0.155
	600	0.531	0.515	0.016	0.157
	525	0.533	0.515	0.018	0.157
	500	0.534	0.517	0.017	0.158
	475	0.536	0.519	0.017	0.160
	425	0.540	0.524	0.016	0.164
	450	0.542	0.525	0.016	0.167
	400	0.553	0.538	0.015	0.170
	375	0.559	0.543	0.016	0.176

Appendix C Character N-gram Features: Full Results

Table C1

Best-performing Experiment Settings for Character Unigrams (n=1) for 700, 1,400, and 2,100 Word Lengths

Sample Word Number	Number of Features	*Cllr*	*Cllr^{min}*	*Cllr^{cal}*	*EER*
700	50	0.807	0.796	0.011	0.286
	60	0.811	0.798	0.013	0.284
	70	0.813	0.801	0.012	0.293
	90	0.814	0.797	0.017	0.291
	80	0.815	0.801	0.014	0.290
	40	0.819	0.807	0.011	0.293
	30	0.838	0.827	0.011	0.301
	20	0.952	0.943	0.009	0.392
	10	0.968	0.959	0.009	0.417
	5	0.988	0.980	0.008	0.441
1,400	90	0.656	0.644	0.012	0.210
	70	0.661	0.648	0.013	0.208
	80	0.661	0.649	0.013	0.210
	50	0.662	0.651	0.011	0.211
	60	0.663	0.651	0.012	0.212
	40	0.679	0.667	0.013	0.218
	30	0.712	0.702	0.010	0.235
	20	0.879	0.869	0.010	0.331
	10	0.919	0.912	0.007	0.357
	5	0.956	0.948	0.008	0.389
2,100	60	0.553	0.538	0.015	0.169
	90	0.553	0.541	0.012	0.165
	50	0.554	0.542	0.012	0.170
	70	0.555	0.542	0.013	0.166
	80	0.557	0.547	0.011	0.168
	40	0.566	0.554	0.012	0.179
	30	0.609	0.596	0.013	0.200
	20	0.787	0.777	0.010	0.279
	10	0.867	0.859	0.008	0.320
	5	0.938	0.925	0.013	0.374

Table C2

Best-performing Experiment Settings for Character Bigrams (n=2) for 700, 1,400, and 2,100 Word Lengths

Sample Word Number	Number of Features	*Cllr*	*Cllr^{min}*	*Cllr^{cal}*	*EER*
700	1000	0.722	0.710	0.012	0.240
	875	0.722	0.708	0.014	0.239
	825	0.723	0.710	0.013	0.243
	975	0.724	0.710	0.014	0.241
	850	0.724	0.711	0.013	0.242
	900	0.724	0.711	0.013	0.243
	950	0.725	0.711	0.014	0.243
	925	0.726	0.710	0.015	0.242
	800	0.726	0.713	0.014	0.246
	775	0.729	0.714	0.015	0.245
1,400	1000	0.498	0.483	0.015	0.150
	875	0.500	0.487	0.014	0.150
	700	0.501	0.487	0.015	0.152
	975	0.501	0.489	0.013	0.154
	900	0.501	0.487	0.014	0.152
	950	0.502	0.488	0.014	0.153
	850	0.502	0.489	0.013	0.153
	675	0.503	0.491	0.011	0.153
	825	0.503	0.489	0.014	0.153
	725	0.504	0.490	0.013	0.152
2,100	1000	0.356	0.342	0.014	0.101
	875	0.356	0.344	0.012	0.104
	850	0.357	0.344	0.013	0.105
	950	0.357	0.342	0.015	0.101
	900	0.358	0.347	0.012	0.102
	975	0.358	0.343	0.015	0.103
	925	0.358	0.344	0.015	0.101
	825	0.359	0.346	0.013	0.105
	675	0.360	0.349	0.010	0.103
	700	0.362	0.350	0.012	0.106

Table C3

Best-performing Experiment Settings for Character Trigrams (n=3) for 700, 1,400, and 2,100 Word Lengths

Sample Word Number	Number of Features	*Cllr*	*Cllr*^{min}	*Cllr*^{cal}	*EER*
700	925	0.817	0.797	0.019	0.287
	900	0.817	0.799	0.018	0.288
	950	0.819	0.796	0.023	0.288
	875	0.819	0.800	0.019	0.288
	850	0.820	0.801	0.019	0.287
	1000	0.821	0.802	0.019	0.287
	975	0.821	0.801	0.021	0.285
	825	0.822	0.805	0.017	0.291
	800	0.823	0.806	0.017	0.291
	550	0.823	0.807	0.016	0.299
1,400	950	0.580	0.567	0.013	0.182
	975	0.581	0.566	0.015	0.181
	1000	0.583	0.566	0.017	0.183
	925	0.583	0.571	0.013	0.184
	875	0.584	0.571	0.014	0.186
	900	0.585	0.572	0.013	0.186
	850	0.590	0.578	0.012	0.186
	825	0.593	0.582	0.012	0.187
	800	0.598	0.586	0.012	0.189
	750	0.601	0.587	0.014	0.191
2,100	950	0.429	0.418	0.012	0.128
	975	0.432	0.420	0.012	0.128
	1000	0.432	0.420	0.013	0.128
	925	0.436	0.424	0.011	0.130
	875	0.437	0.425	0.012	0.134
	900	0.438	0.428	0.011	0.131
	850	0.446	0.436	0.010	0.134
	825	0.449	0.438	0.011	0.136
	800	0.452	0.442	0.010	0.137
	775	0.457	0.444	0.013	0.136

Table C4

Best-performing Experiment Settings for Character Trigrams (n=3) for 700, 1,400, and 2,100 Word Lengths

Sample Word Number	Number of Features	*Cllr*	*Cllr*^{min}	*Cllr*^{cal}	*EER*
700	1000	0.784	0.769	0.014	0.274
	975	0.789	0.772	0.018	0.275
	900	0.792	0.774	0.018	0.276
	950	0.793	0.774	0.018	0.277
	925	0.793	0.777	0.016	0.276
	875	0.796	0.777	0.019	0.276
	825	0.796	0.777	0.019	0.277
	850	0.796	0.777	0.020	0.275
	800	0.798	0.780	0.018	0.278
	775	0.801	0.784	0.017	0.279
1,400	1000	0.531	0.516	0.015	0.160
	825	0.535	0.520	0.015	0.166
	900	0.536	0.522	0.014	0.165
	800	0.536	0.522	0.014	0.167
	925	0.536	0.522	0.015	0.164
	975	0.537	0.522	0.015	0.163
	950	0.537	0.521	0.016	0.165
	850	0.539	0.524	0.015	0.166
	875	0.539	0.523	0.016	0.168
	775	0.542	0.528	0.014	0.167
2,100	1000	0.387	0.375	0.012	0.111
	800	0.388	0.377	0.011	0.114
	825	0.389	0.377	0.012	0.113
	975	0.392	0.377	0.014	0.111
	775	0.392	0.380	0.012	0.114
	900	0.392	0.379	0.013	0.113
	875	0.393	0.382	0.011	0.114
	850	0.394	0.382	0.012	0.115
	950	0.395	0.381	0.014	0.111
	925	0.396	0.381	0.014	0.111

Appendix D Part-of-speech N-gram Features: Full Results

Table D1

Best-performing Experiment Settings for Part-of-speech Unigrams (n=1) for 700, 1,400, and 2,100 Word Lengths

Sample Word Number	Number of Features	*Cllr*	*Cllr^{min}*	*Cllr^{cal}*	*EER*
700	40	0.752	0.741	0.011	0.259
	30	0.758	0.743	0.015	0.261
	20	0.809	0.797	0.012	0.295
	10	0.862	0.848	0.013	0.322
	5	0.938	0.926	0.012	0.377
1,400	40	0.564	0.549	0.015	0.173
	30	0.573	0.556	0.016	0.171
	20	0.645	0.632	0.013	0.204
	10	0.732	0.721	0.011	0.248
	5	0.860	0.837	0.023	0.307
2,100	30	0.439	0.425	0.014	0.129
	40	0.440	0.427	0.013	0.128
	20	0.522	0.510	0.012	0.158
	10	0.630	0.619	0.011	0.199
	5	0.792	0.772	0.020	0.265

Table D2

Best-performing Experiment Settings for Part-of-speech Bigrams (n=2) for 700, 1,400, and 2,100 Word Lengths

Sample Word Number	Number of Features	*Cllr*	*Cllr^min^*	*Cllr^cal^*	*EER*
700	525	0.714	0.696	0.018	0.237
	450	0.715	0.698	0.017	0.234
	475	0.715	0.700	0.016	0.238
	425	0.717	0.696	0.021	0.230
	500	0.717	0.701	0.016	0.239
	550	0.718	0.700	0.017	0.238
	400	0.719	0.697	0.023	0.234
	600	0.721	0.708	0.014	0.237
	575	0.722	0.705	0.017	0.237
	375	0.725	0.703	0.022	0.238
1,400	400	0.486	0.473	0.013	0.144
	425	0.487	0.474	0.013	0.147
	525	0.488	0.474	0.013	0.146
	600	0.488	0.476	0.012	0.146
	500	0.489	0.477	0.012	0.146
	450	0.489	0.478	0.011	0.148
	475	0.489	0.476	0.013	0.148
	550	0.489	0.476	0.013	0.146
	375	0.490	0.478	0.012	0.145
	575	0.491	0.479	0.012	0.146
2,100	400	0.342	0.329	0.012	0.093
	425	0.343	0.331	0.013	0.093
	525	0.344	0.331	0.012	0.094
	475	0.345	0.332	0.013	0.094
	450	0.347	0.334	0.013	0.094
	500	0.348	0.334	0.014	0.093
	550	0.350	0.335	0.015	0.095
	600	0.350	0.336	0.014	0.097
	375	0.351	0.340	0.011	0.095
	575	0.355	0.340	0.014	0.099

Table D3

Best-performing Experiment Settings for Part-of-speech Trigrams (n=3) for 700, 1,400, and 2,100 Word Lengths

Sample Word Number	Number of Features	*Cllr*	*Cllr^{min}*	*Cllr^{cal}*	*EER*
700	600	0.764	0.743	0.020	0.256
	575	0.764	0.740	0.024	0.255
	550	0.767	0.744	0.023	0.256
	500	0.771	0.750	0.021	0.257
	525	0.771	0.746	0.025	0.254
	375	0.773	0.760	0.013	0.269
	325	0.774	0.759	0.015	0.266
	475	0.775	0.752	0.024	0.256
	300	0.777	0.764	0.013	0.264
	450	0.781	0.757	0.024	0.258
1,400	600	0.535	0.516	0.018	0.162
	575	0.536	0.517	0.020	0.163
	550	0.539	0.519	0.020	0.163
	525	0.546	0.528	0.018	0.167
	500	0.547	0.530	0.017	0.167
	475	0.553	0.537	0.016	0.170
	425	0.554	0.543	0.011	0.177
	450	0.556	0.542	0.014	0.174
	400	0.557	0.546	0.011	0.176
	375	0.560	0.547	0.013	0.179
2,100	600	0.399	0.387	0.012	0.118
	575	0.406	0.394	0.012	0.121
	500	0.408	0.393	0.015	0.128
	550	0.408	0.395	0.013	0.123
	525	0.409	0.397	0.012	0.127
	475	0.417	0.404	0.013	0.131
	450	0.419	0.407	0.012	0.132
	425	0.424	0.413	0.011	0.130
	400	0.428	0.416	0.012	0.133
	375	0.434	0.418	0.016	0.130