Neural and Non-neural Approaches to Authorship Attribution

A Thesis submitted to the University of Sheffield
for the degree of Doctor of Philosophy in the Faculty of Engineering

by

Yunita Sari

Department of Computer Science
The University of Sheffield

September 2018

Neural and Non-neural Approaches to Authorship Attribution

ABSTRACT

This thesis explores a range of authorship attribution approaches and proposes new techniques to improve performance. Authorship attribution is the task of identifying the author of a text. It has attracted attention due to its relevance to a wide range of applications including forensic investigation and plagiarism detection. An array of features and approaches have been applied to this task. However, there has been a lack of study which involves multiple datasets or uses a range of different classifiers. Therefore, in this thesis we explore both neural and non-neural network models and use different feature representations on multiple datasets.

We begin with a short introduction to authorship attribution in Chapter 1. A more comprehensive review of authorship attribution and its related tasks is given in Chapter 2. In Chapter 3 we introduces a novel analysis using topic modeling to examine the conditions under which each type of authorship attribution feature is useful. Chapter 4 explores the implementation of language modeling for authorship attribution. We describe the feature selection issue in standard authorship attribution approaches and evaluate whether n-gram language modeling can help to address the problem. Furthermore, we implement A Long Short Term Memory (LSTM) language model for authorship attribution and assess its effectiveness for the task.

In Chapter 5 we present our work on using continuous representations for authorship attribution. In contrast to previous work, which uses discrete feature representations, our model learns continuous representations for n-gram features via a neural network jointly with the classification layer. The proposed model outperforms the state-of-the-art on two datasets, while producing comparable results on the remaining two. In addition, we describe our novel extension of the proposed models and show how the analysis in Chapter 3 helps to improve the attribution accuracy. Finally, we demonstrate how the authors' demographic profiles can help improve task performance via Multi Task Learning (MTL). In Chapter 6 we highlight the contributions of this thesis and propose directions for future research in this area.

Acknowledgements

First of all, I would like to thank my supervisors Mark Stevenson and Andreas Vlachos for their continuous support and guidance through my PhD. Mark and Andreas provided me with invaluable advice, help, and fruitful discussion. They shaped me as a researcher and made the PhD journey more enjoyable. I truly appreciate all the time and effort they took for me. I am very grateful to my PhD panel committee members, Lucia Specia and Gordon Fraser, for their feedback and constructive comments. My examiners, Hamish Cunningham and Michael Oakes greatly helped me in improving this thesis. I would like to express my gratitude to the Indonesia Endowment Fund for Education (LPDP) which funded my studies in Sheffield.

I would also like to thank all of my colleagues in the Natural Language Processing group, in particular everybody in room G28. I really appreciate their helpful discussions and their kindness. I want to thank all the Indonesians in Indosoc and KIBAR Sheffield for all the fun, laughter, and delicious food. Special thanks go to my ex-house mates in 194 Oxford Street for listening to me patiently and being happy testers (my assumption) for my experimental cakes. Big thanks to Bu Dita, my sixth supervisor, for chats and advices which keep me sane until the end of the PhD. Finally, I want to thank my family, Bapak, Ibu, Mas Ipang, Mba Ani, Mba Tyas, Salsa and Lutfi for their unreserved support and love. For Bapak, I have fulfilled my promise to finish this journey no matter what.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Swinson and Reyna (2013) described authorship attribution simply as “the task of identifying the author of a text”. This field is part of *stylometry*, the study of style which was originally applied to handwritten texts. In more recent work, stylometry has been widely applied to digital texts and computer code (Juola, 2008; Stamatatos, 2009a; Caliskan-Islam et al., 2015; Shrestha et al., 2017). The famous case of the Federalist Papers can be considered as an early example of authorship attribution (Juola, 2008). The Federalist Papers are a set of newspaper essays published in the late 18$^{\text{th}}$ century by an anonymous author named ‘Publius’. Those papers were published to persuade New York residents to approve the newly proposed Constitution of the United States. Lately, it has become known that those 85 essays were written by three authors: John Jay, Alexander Hamilton, and James Madison. Five essays were written by Jay, 14 essays by Madison and 51 essays by Hamilton. There are also 12 disputed essays which have been claimed by both Madison and Hamilton.

Authorship attribution has great potential to be used in various authorship analysis applications; such as history and literary science (e.g. determining the author of a disputed or anonymous text document) (Klarreich, 2003; Oakes, 2004; Burrows, 2002; Hoover, 2004), forensic investigation (e.g. identifying authors in anonymous or phishing email) (Chaski, 2005; Grant, 2007; Iqbal et al., 2010; Lambers and Veenman, 2009; Gollub et al., 2013), plagiarism (e.g. detecting collaboration in the document) (Gollub et al., 2013; Kimler, 2003) or even used as evidence in courts of law (Morton and Michaelson, 1990). An example of a real case is the role of authorship attribution in revealing J. K. Rowling who used the pen name Robert Galbraith for her book “The Cuckoo’s Calling” (Zimmer, 2013). By using four different linguistic variables such as the distribution of word lengths and the frequency of 100 most common words, the analysis pointed strongly to Rowling as the true author.

Most work in the area of authorship attribution tries to determine whether an individual can be distinguished based on their writing style. Halteren et al. (2005) proposed the idea of a "human stylome" as "specific properties of writing style that can be used to identify the author." Similarly, Juola (2008) used the term "authorial fingerprint" to describe "characteristic pattern of language used in people's writings." There was an idea to use a person's writing style as identification key like a fingerprint. However, it has yet to be determined whether there are specific writing style features that can identify a person, like a fingerprint (Malyutov, 2006). These writing style features need to be constant regardless of aspects that can vary such as the writing time and the genre of the document.

The study of authorship attribution has been progressing quickly and has been extended into more advanced problems such as authorship verification and adversarial stylometry (see details in Section 2.3). The majority of existing authorship attribution approaches apply supervised machine learning algorithms with a wide range of features. Most of those works focused on feature engineering by exploring potential features which could improve the attribution performance (see Section 2.2). Function words and character *n*-grams are two features that have been proved to be effective for capturing an author's writing style (Mosteller and Wallace, 1964; Peng et al., 2003; Argamon and Levitan, 2005; Koppel et al., 2005; Juola and Baayen, 2005; Zhao and Zobel, 2005; Stamatatos, 2013; Schwartz et al., 2013). On the other hand, some approaches have explored different architectures for the problem. Rather than using a standard machine learning algorithm such as a Support Vector Machines (SVMs), they utilised neural network-based models (Bagnall, 2015; Sari et al., 2017; Shrestha et al., 2017).

A range of approaches have been applied for authorship attribution with different features and architectures. However, they have not been systematically compared. The lack of continuity in authorship attribution research causes difficulties to identify which approaches are most reliable in any particular circumstance. Evaluation forums such as PAN (see Section 2.4) help to address this problem. Yet, their small size of released datasets prevent the participants from exploring more advanced techniques. Juola (2008) emphasised that accuracy is not the only consideration in authorship attribution. An authorship attribution system is expected to be adaptive to the language, genre, size, of the available documents. In addition, authorship attribution is a task with cross-disciplinary interests. A good research work needs to be validated in various areas and problems. Thus, the expected levels of accuracy and the circumstances where it might be expected to drop can be easily defined.

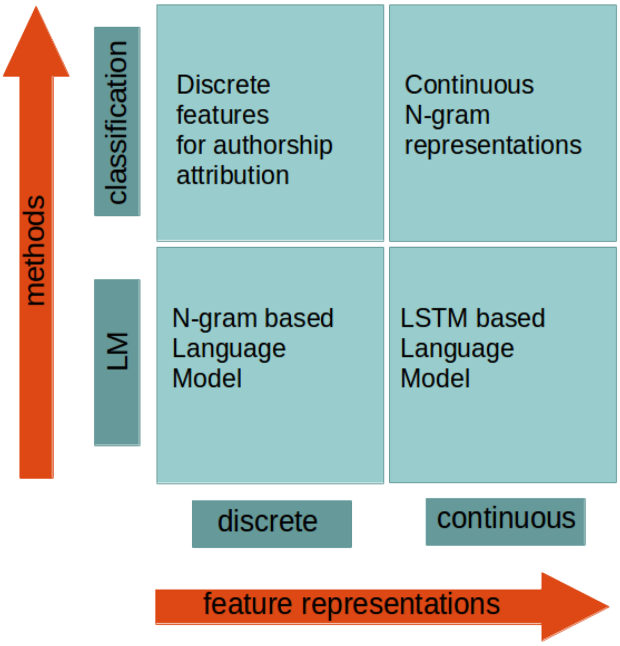

methods
classification
LM
Discrete features for authorship attribution
Continuous N-gram representations
N-gram based Language Model
LSTM based Language Model
discrete
continuous
feature representations

Figure 1.1: Main research objective

This thesis aims to provide more clear direction of the authorship attribution approaches by exploring four different techniques as shown in Figure 1.1. In general, the approaches can be divided along two dimensions: method and representation. Discrete and continuous representations are combined with two types of methods: Language Model (LM) and supervised classification. In continuous representations, each feature is represented as a d-dimensional real-valued vector and its values are learned via a neural-network based model. In this way, similar features are likely to have similar vectors (Goldberg and Hirst, 2017). In contrast, in discrete representations, features are completely independent from one another. Both non-neural and neural network-based approaches are also explored in this thesis. There are five datasets involved in our studies including Judgment (Seroussi et al., 2011), CCAT10, CCAT50 (Stamatatos, 2008), IMDb62 (Seroussi et al., 2010) and The Blog Authorship Corpus (Schler et al., 2006). These datasets are commonly used in previous literature and represent a range of characteristics in terms of the number of authors, topic/genre and document length. By conducting exploration on different approaches, we aim to address the following problems:

- Previous work has explored an extensive array of authorship attribution features (see Section 2.2). However, there has been a lack of analysis of the behavior of features across multiple datasets or using a range of classifiers. Consequently, it is difficult to determine which types of features will be most useful for a particular authorship attribution dataset. This thesis explores how the characteristics of an authorship attribution dataset affect the usefulness of different types of features.

- Feature selection is an important step in authorship attribution which usually involves setting threshold to remove uninformative features (Scott and Matwin, 1999). However, defining an optimal threshold can be problematic, because rare features may contain essential information about the author's writing style. Peng et al. (2003) avoided the problem by implementing n-gram based language models (LM). The model enables to include every feature without experiencing sparse data problems. Their approach obtained performance higher than 90% in accuracy. However they conducted experiments on a limited type of datasets which may not reflect the effectiveness of the models. This work examines to what extent the n-gram language model can benefit the authorship attribution task in various types of datasets.

- One of the major drawbacks of an n-gram-based language model is it usually depends only on the previous two or three words. This limits the model from capturing information from a longer context which might be useful for authorship attribution. The above problems can theoretically be solved using the Long Short Term Memory (LSTM)-based language model (Sundermeyer et al., 2012). This thesis investigates whether the application of an LSTM-based language model can help to improve authorship attribution performance.

- In authorship attribution, features are commonly represented in discrete form. However, this representation suffers from data sparsity and does not consider the semantic relatedness between features. For example, in the bag-of-words representation, the words "Paris" and "London" are not connected. In contrast, using a continuous representation, the distance is closer since those words are semantically similar. Continuous representations have been shown to be helpful in a wide range of tasks in natural language processing (Mikolov et al., 2013b; Bansal et al., 2014; Joulin et al., 2017; Li et al., 2016; Rahimi et al., 2017). This thesis explores continuous n-gram representations for authorship attribution tasks.

- Previous work (Hovy, 2015) demonstrated that the author's demographic profile can help to improve the performance of text classification tasks such as sentiment analysis and topic detection via a Multi Task Learning (MTL) framework. No previous work on authorship attribution has tried to implement this approach, despite it being promising. This thesis examines the role of age and gender information to improve authorship attribution performance.

1.1 Contributions

This thesis makes the following research contributions:

- Proposes a novel analysis using topic modeling to examine the conditions under which each type of authorship attribution feature is useful.

- Explores the application of n-gram based language model for authorship attribution.

- Investigates whether an LSTM-based language model can help to improve the attribution performance.

- Proposes the implementation of continuous n-gram representations to address the discrete representations problem.

- Proposes the use of Multi Task Learning (MTL) to jointly learn authorship attribution with gender and age identifications.

- Presents a comparative study of four different authorship attribution approaches representing a range of feature representations (discrete and continuous) and approaches (non-neural and neural network based model).

1.2 Thesis Overview

The remainder of this thesis is structured as follows:

Chapter 2 (Background) provides a review of previous work of authorship attribution. The chapter starts with a review of the current state of authorship attribution. It then presents features and methods that have been used to address this task. Various forms of authorship attribution are discussed including author profiling, open and closed set authorship attribution, adversarial stylometry and authorship verification. In addition, progress in authorship attribution shared tasks is described.

Chapter 3 (Exploring the Most Useful Features for Authorship Attribution) introduces a novel analysis using topic modeling to examine the conditions under which each type of authorship attribution feature is useful. In addition, previous work which utilized the same datasets involved in our experiments is also presented.

Chapter 4 (Language Model for Authorship Attribution) describes feature selection problems in authorship attribution and how n-gram language modeling-based approaches can help to address them. Among existing authorship attribution methods, this model has rarely been explored. A separate language model is created for each of the authors. The author of an unseen document is determined by comparing the document against each model and choosing the one with lowest perplexity. Results indicate that the feature selection problems can be addressed. However, the model failed to outperform the previous results due to the problem of data sparsity. Furthermore, in this chapter we investigate the effectiveness of a neural network model by implementing LSTM-based language modeling for authorship attribution. Unlike the n-gram language model, the recurrent connections of LSTM allows to capture information from long context sequences. However, the experimental results show that the information does not benefit the authorship attribution performance. A thorough discussion is presented on this case.

Chapter 5 (Continuous *N*-gram Representations for Authorship Attribution) presents work on using continuous representations for authorship attribution. In contrast to most previous work which uses discrete representations, this model learns continuous representations for n-grams via a neural network jointly with the classification layer. Experimental results demonstrate that the proposed model outperforms the state-of-the-art on two datasets, while producing comparable results on the remaining two. In addition, we describe our novel extension of the proposed models and show how the analysis in Chapter 3 helps to improve the attribution accuracy. Finally, we propose Multi Task Learning (MTL) which jointly learns authorship attribution with gender and age identifications. Results from the experiments show a consistent improvement on the performance along the increase of the author numbers.

Chapter 6 (Conclusions and Future Work) provides the conclusion of this thesis and presents future directions for research in this area.

1.3 Published Material

The following publications are related to the work reported in this thesis:

1. Sari, Y., Stevenson, M., and Vlachos, A. (2018). Topic or Style? Exploring the Most Useful Features for Authorship Attribution. Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018), Santa Fe, USA.

2. Sari, Y., Vlachos, A., and Stevenson, M. (2017). Continuous n-gram representations for authorship attribution. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 267– 273, Valencia, Spain. Association for Computational Linguistics.

3. Sari, Y. and Stevenson, M. (2016). Exploring Word Embeddings and Character n-grams for Author Clustering—Notebook for PAN at CLEF 2016. In Balog, K., Cappellato, L., Ferro, N., and Macdonald, C., editors, CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, Evora, Portugal. CEUR-WS.org.

4. Sari, Y. and Stevenson, M. (2015). A Machine Learning-based Intrinsic Method for Cross-topic and Cross-genre Authorship Verification—Notebook for PAN at CLEF 2015. In Cappellato, L., Ferro, N., Jones, G., and San Juan, E., editors, CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France. CEUR-WS.org.

5. Sari, Y. (2015). Gender Identification for Adversarial Writing. In Proceedings of the ESSLLI 2015 Student Session, Barcelona, Spain.

Chapter 2

Background

This chapter presents background related to authorship attribution tasks. It gives a detail description of various approaches and features that will be used in the experiments presented in later chapters. In addition, it present different types of tasks related to authorship attribution such as author profiling and verification. We also describe an evaluation forum which covers several authorship identification shared tasks and show how it helps to accelerate the progress in authorship attribution field.

First, we begin this chapter by presenting a short overview of current research state of authorship attribution in Section 2.1. It covers description of methods that have been employed for this task. We then describe various types of authorship attribution features in Section 2.2. In Section 2.3 we discuss different types of tasks related to authorship attribution. Finally, the overview of authorship attribution evaluation forums are provided in Section 2.4 followed by a summary in Section 2.5

2.1 State of The Authorship Attribution Task

From a machine learning perspective, the authorship attribution task can be treated as a form of text classification. Let $D = d_1, d_2, ..., d_n$ be a set of documents and $A = a_1, a_2, ..., a_m$ a fixed set of candidate authors. The task of authorship attribution is to assign an author to each of the documents in D. The challenge in authorship attribution is that identifying the topic preference of each author is not sufficient; it is necessary to also capture their writing style (Stamatatos, 2013). This task is more difficult than determining the topic of a text, which is possible by identifying domain-indicative lexical items since writing style cannot be fully captured by an author's choice of vocabulary.

Rudman (1997), Juola (2008), Stamatatos (2009a), and Gollub et al. (2013) provided comprehensive reviews of the state-of-the-art in authorship attribution work describing the current state of the field at the time of writing. All four agreed that research in authorship attribution started in the late 19th century with the work of Mendenhall (1887). In his research, Mendenhall proposed a simple approach for author attribution of texts belong to Bacon, Marlowe, and Shakespeare using simple stylometric features. A curve was created for each text document expressing a link between word length and its frequency; which later could be used as the basis to determine the author of the text. Later studies (e.g. Smith (1983)), proved that Mendenhall's proposed method is unreliable for authorship attribution.

Statistical method started to be used in the mid 20th century (Zipf, 1933; Yule, 1939, 1944) . Koppel et al. (2009) categorized that early work as part of the *unitary invariant approach*, in which authors are discriminated by only a single numeric function. Attribution studies that consider this approach include Brinegar (1963), Foster (1989) (word length); Williams (1940), Yule (1939) (sentence length); and Holmes (1992), Yule (1944) (vocabulary richness). However, Hoover (2003) and Grieve (2007) shared a similar opinion that unitary invariant approach was not reliable enough to identify authors. Grieve conducted experiments which involved thirty-nine different types of textual measurements commonly used in attribution studies. His experiments which were performed using an identical attribution algorithm and tested on the same dataset proved that combination of the word and punctuation mark profiles are the best feature sets among others. He argued that better performance can be obtained by using larger number of textual measurement.

More recent attribution work started to apply the *multivariate analysis approach*, in which multiple stylistic features are combined, and then a statistical multivariate discriminant analysis applied. Mosteller & Wallace (1964) implemented a multivariate analysis approach to the Federalist Papers by using 30 function words (e.g. the, of, about, and, etc.) frequency as the feature and Naïve Bayes as the classifier. For early attribution studies, the Federalist Papers were considered as the ideal testing ground as there is a well defined set of candidate authors, sets of known authorship for all the candidate authors, and a set of texts of disputed authorship. In addition, the papers have the same genre and thematic area (Stamatatos, 2009a). Mosteller & Wallace proved that high-frequency function words are effective features to discriminate authors. Because of this promising result, their work has been marked as the most successful attribution work for that period.

Another approach which can be considered as the most recent and widely used method is *supervised machine learning based approaches*. In this technique, the machine learning algorithms learn to find the boundaries between classes that minimize the classification cost function. The algorithms create the classification models based on training texts which are represented as labeled numerical vectors. The models then can be used to assign classes to the unlabeled test documents. Figure 2.1 illustrates supervised machine learning applied for authorship attribution task which is divided into two phases: training and test. Among machine learning algorithms, Support Vector Machine (SVM) (Cortes and Vapnik, 1995) has been proved to be an effective method. The ability of SVM to handle sparse and high-dimensional data, make it suitable for authorship attribution and in general for the text classification task (Stamatatos, 2013).

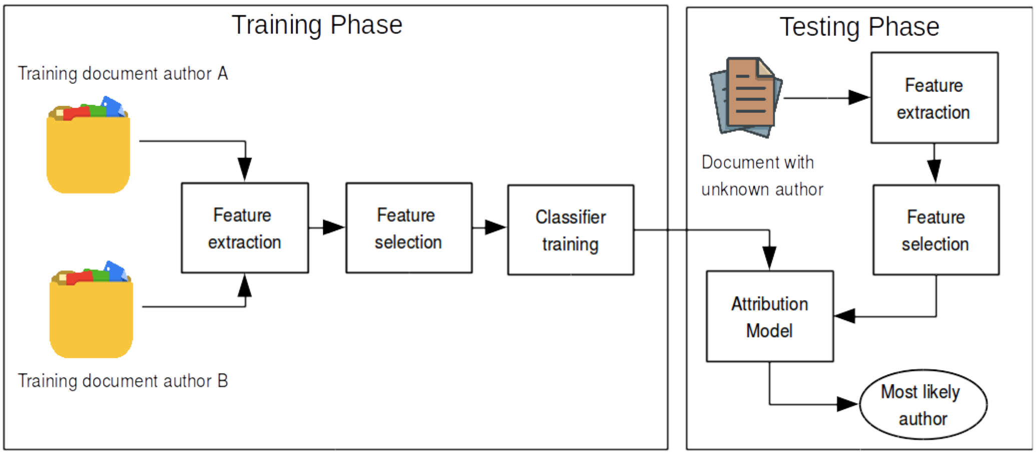


Figure 2.1: Authorship attribution with supervised machine learning approach

In addition to SVM, neural network based methods recently have enjoyed a resurgence in popularity. Some authorship attribution work which has tried to implement this approach reported an improvement in the results (Bagnall, 2015; Sari et al., 2017; Shrestha et al., 2017). For example Shrestha et al. (2017) implemented a character level Convolutional Neural Network (CNN) for short text. The authors argued that the architecture of CNN which consists of convolutional and pooling layers is suitable to capture local interactions between characters. This information is then aggregated to learn high-level patterns for modeling the authors' writing style. Evaluated on a twitter dataset (Schwartz et al., 2013) with 1,000 tweets per author, the CNN model successfully gained improvement over the previous work with 76.1% accuracy. However, as expected from the neural-based method which usually needs a large number of training data, the

CNN performance decreased significantly when less training data was provided. Another work by Sari et al. (2017) proposed a simpler neural-based architecture for authorship attribution. Their model adopted the feed-forward neural network architecture of fastText (Joulin et al., 2017). With a simpler architecture, the model outperformed previous work on the CCAT50 dataset (Stamatatos, 2008) with only 50 training documents per author.

2.2 Authorship Attribution Features

Feature engineering is one of the most important stages in the development of authorship attribution tasks. A large amount of work has explored different types of features (Grieve, 2007; Guthrie, 2008; Brennan, 2012; Sapkota et al., 2015). Stamatatos (2009a) listed the basic authorship attribution features and the tools to measure them (see Table 2.1). Authorship attribution features are often referred to the term *stylometric features* due to the initial goal of the authorship attribution task which more focused on exploiting the authors' writing style. However, there is evidence that topical information might also be useful (Koppel et al., 2009). Nevertheless, the feature is not the only factor which determines the accuracy of authorship attribution. Attribution techniques, feature selection and extraction technique and the nature of the corpus certainly will also contribute to the performance (Forsyth and Holmes, 1996; Houvardas and Stamatatos, 2006; Koppel et al., 2006; Luyckx and Daelemans, 2010; Savoy, 2013; Stamatatos, 2013). In this section details of most common stylometric features including lexical, character, syntactic semantic, and bag-of-words features are discussed. In addition, a range of feature selection techniques is also presented.

2.2.1 Lexical Features

Lexical features are commonly used in authorship attribution work as they provide rich information about the author's writing style. In addition, most lexical features can be applied to any language[1] and corpus with no additional requirement (Stamatatos, 2009a). Some examples of simple lexical features include word frequencies, word n-grams, function words, function word n-grams, hapax legomena[2], morphological information (lemma, stem, case, mood, etc.), word, sentence and paragraph length, grammatical errors and slang words. These features have

[1] Except for certain natural language, e.g. Chinese which has a less-defined concept of sentence (Huang and Chen, 2011)

[2] Word that only appeared once in the document

Feature type	**Description**	**Required tools and resource**
Lexical	token based (word length, sentence length, etc)	Tokenizer [sentence splitter]
	Vocabulary richness	Tokenizer
	Word Frequencies	Tokenizer
	Word n-grams	Tokenizer, [stemmer, lemmatizer]
	Errors	Tokenizer, orthographic, spell checker
Character	Character types (letter, digits, etc)	Character dictionary
	Character n-grams(fixed length)	-
	Character n-grams(variable length)	Feature selector
	Compression	Text compression tools
Syntactic	Part of speech	Tokenizer, sentence splitter, POS tagger
	Chunks	Tokenizer, sentence splitter, POS tagger, text chunker
	Sentence and phrase structure	Tokenizer, sentence splitter, POS tagger, text chunker, partial parser
	Rewrite rules frequencies	Tokenizer, sentence splitter, POS tagger, text chunker, full parser
	Errors	Tokenizer, sentence splitter, syntactic spell checker
Semantic	Synonyms	Tokenizer, POS tagger, thesaurus
	Semantic Dependencies	Tokenizer, sentence splitter, POS tagger, text chunker, partial parser, semantic parser
	Functional	Tokenizer, sentence splitter, POS tagger, specialized dictionaries
Appl. Specific	Structural	HTML parser, specialized parser
	Content Specific	Tokenizer, stemmer,lemmatizer, specialized dictionaries
	Language Specific	Tokenizer, stemmer, lemmatizer, specialized dictionaries

Table 2.1: Description of authorship attribution features (Stamatatos, 2009a)

been widely used since early attribution works which utilized only simple statistical approach (Mendenhall, 1887) to more recent work that applied complex machine learning techniques (Koppel et al., 2009; Seroussi et al., 2013). Several lexical features are found to be ineffective when used alone (Burrows, 1992; Grieve, 2007). Grieve evaluated a range of features and found that word length, sentence length, and vocabulary richness appeared to be of little use to distinguish authorship. On the other hand, he observed that word frequency, punctuation mark frequency, and *n*-gram frequency are effective for distinguishing among a small number of authors with accuracy more than 90%.

In content-based text classification representing documents as vectors of words frequencies is the most common approach (the bag-of-words approach). In this approach, common words like articles, prepositions, pronouns, etc., (known as function words) will often be discarded as being unhelpful. In contrast to content-based technique, authorship attribution utilizes those common words to capture the author's writing style. Function words have been proved to be effective features because authors can not consciously control the usage of those words in their writing. Several studies have reported the efficacy of function words in authorship attribution including Mosteller and Wallace (1964), Argamon and Levitan (2005), Koppel et al. (2005), Juola and Baayen (2005), Zhao and Zobel (2005).

2.2.2 Character Features

Grieve (2007) reported that character features were effective in capturing stylistic information. Along with lexical features, character features have become one of the popular features used in many author attribution tasks since they can be easily extracted from text (Stamatatos, 2009a). Commonly used character features include letter frequencies, alphabetic characters count, uppercase and lowercase characters count, digit count, punctuation mark count, and character *n*-grams (de Vel et al., 2001; Zheng et al., 2006).

Character *n*-grams have been reported to outperform other character features (Grieve, 2007). Other studies also reported the successful application of this approach (Kjell, 1994; Forsyth and Holmes, 1996; Stamatatos, 2006; Peng et al., 2003; Keselj et al., 2003; Juola, 2004). Beside being easily available and effectively capturing stylistic information, character *n*-grams are also tolerant to noise (Stamatatos, 2009a). Character *n*-gram representations are not significantly affected by spelling errors or strange use of punctuation marks in the text. An example is given by Stamatatos: the words 'simplistic' and 'simpilstc' would be considered two different words in word *n*-grams representation, but a character *n*-gram representation would generate many common character *n*-grams.

In addition, character n-grams can be used to capture regular grammatical or orthographic errors which in some cases could represent the character of the authors (Koppel and Schler, 2003).

Other studies used character features with a different approach. Instead of using machine learning methods that usually need training data, compression-based approaches were used (Benedetto et al., 2002; Khmelev and Teahan, 2003; Marton et al., 2005; Oliveira et al., 2013). The archive size of the text file after compression is compared with distance indicating the similarity between documents.

The reason for the success of character n-grams is not well understood. Koppel et al. (2011) argued that the effectiveness comes from the ability of character n-grams to capture both content and stylistic information in the text. Similar conclusions were reported by Sapkota et al. (2015) who analyzed subgroups of character n-grams which have been claimed to represent some linguistic aspects like morphosyntax, thematic content, and style. The predictiveness of each subgroup was evaluated in single and cross-domain settings. The results of their study demonstrate that affixes and punctuation n-grams make a significant contribution towards the effectiveness of character n-grams.

2.2.3 Syntactic Features

A different type of feature set is based on syntactic information produced by analysis tools such as text chunkers and parsers. A number of studies have found that the use of syntactic information could result in better accuracy (Baayen et al., 1996; Chaski, 2005; Gamon, 2004; Hirst and Feiguina, 2007; Stamatatos et al., 2001; van Halteren, 2004). However, the performance for authorship attribution is highly dependent on the accuracy of the syntactic analysis tools. Poor accuracy of the tools will produce noisy features for the attribution classifier.

Many studies used syntactic features in different forms. As an example, Baayen et al. (1996) created 46,403 rewrite rules expressing part of syntactic analysis and then used the frequency of the rules as the features. Another study by Hirst and Feiguina (2007) utilized the bi-gram frequencies of an ordered stream of syntactic labels to discriminate authors of very short texts. Koppel and Schler (2003) proposed an interesting approach by applying syntactic errors such as sentence fragments, and mismatched tense as the attribution features. The most common and simple approach is to use POS tag frequencies or POS tag n-gram frequencies (Argamon-Engelson et al., 1998; Kukushkina et al., 2001; Koppel and Schler, 2003; Zhao and Zobel, 2007)

2.2.4 Semantic Features

Tools are now available to carry out complicated language understanding tasks such as full syntactic parsing, semantic analysis, sentiment analysis, and pragmatic analysis. This progress has made a positive impact on author attribution research. As an example, Bogdanova and Lazaridou (2014) implemented a combination of several semantic, syntactic and lexical features to perform cross-language (English and Spanish) authorship attribution. Interestingly, they introduced sentiment (frequency of positive and negative words) and emotional features (frequency of basic emotions i.e. anger, joy, fear, etc.) produced by SentiWordNet and WordNet-Affect. Expression of sentiments and emotions are reported to be a possible indicator of an author's writing style and personality (Panicheva et al., 2010). Accuracy of analysis tools will affect the performance of authorship attribution in the same way they do for syntactic features.

2.2.5 Bag-of-Words

As mentioned in Section 2.2.1, bag-of-words approaches are the most common method applied for content-based classification tasks such as sentiment analysis and topic classification. Although authorship attribution mostly focuses on style-based features, there is evidence that content words are also useful for authorship attribution tasks (Koppel et al., 2009). This case mostly occurs when there is diversity in the topics discussed within a dataset so that authors can be distinguished based on their topic preferences. There are several options to construct feature representations: the straightforward way is by taking either the absolute or relative term frequency or just representing the text as binary feature vector based on the term occurrences. Another method is to weight terms e.g. by calculating TfIdf values (Salton and Buckley, 1988; Lee and Liu, 2003).

2.2.6 Feature Selection

Approaches to authorship attribution often use combinations of different types of features. Koppel et al. (2009), Stamatatos (2009a), and Grieve (2007) reported that combination set of features could improve the attribution accuracy. However, combining certain features (such as lexical and syntactical features) will certainly increase the complexity of the text representation and may decrease overall classification performance. Feature selection has been widely applied in text classification and is an effective approach for reducing dimensionality. It helps the classifier to avoid over fitting on the training data, diminish irrelevant features, find the best subset of features and may also indirectly increase classification performance.

Most attribution tasks adopted common feature selection approaches from Information Theory area such as Information Gain and Entropy (Houvardas and Stamatatos, 2006), Odds Ratio (Koppel et al., 2006), Kolmogorov complexity (Juola, 2008) and chi-squared (Grieve, 2007; Luyckx and Daelemans, 2008). However, simple feature selection techniques such as term frequency are still commonly applied (Burrows, 1987, 1992; Hoover, 2003).

Several works (Forsyth and Holmes, 1996; Houvardas and Stamatatos, 2006; Koppel et al., 2006; Savoy, 2013) were conducted to compare the effect of certain feature selection approaches for authorship attribution. Interestingly, frequency-based feature selection, the simplest method, outperformed other feature selection approaches such as Information Gain and Odds Ratio. Savoy (2013) performed a comprehensive comparison study of manual selection with six other selection methods including (df- document frequency, X2 - chi-square, IG- Information Gain, PMI- pointwise mutual information, OR- Odd Ratio and DIA – Darmstadt Indexing Approach). Document frequency (df) as a relatively simple selection method and Information Gain (IG) were found to be the most effective approaches compared to the others by achieving accuracy of 97.7%.

2.3 Authorship attribution Forms

In this section various forms of the authorship attribution task are described. Koppel et al. (2009) and Juola (2008) divide the author attribution problem into three main problem sets as shown in Figure 2.2. The first problem is *closed set authorship attribution* (Section2.3.1). In this problem, the challenge is to determine the author of a piece of text where the set of authors is known. This case is similar to multi-class classification problem where the classifier needs to identify the correct class for a particular entity. The second problem is *author verification* (Section 2.3.2) also known as open set authorship attribution. In this problem, the true author might not be in the candidate set. The main challenge is to verify whether a suspect is or is not the author of a document. The open set problem usually harder than the closed problem.

The third problem is *author profiling.* In this case, we need to provide as much information as possible about the author. The information can be psychological (i.e. author personality, mental health, native speaker/not), sociological (e.g.

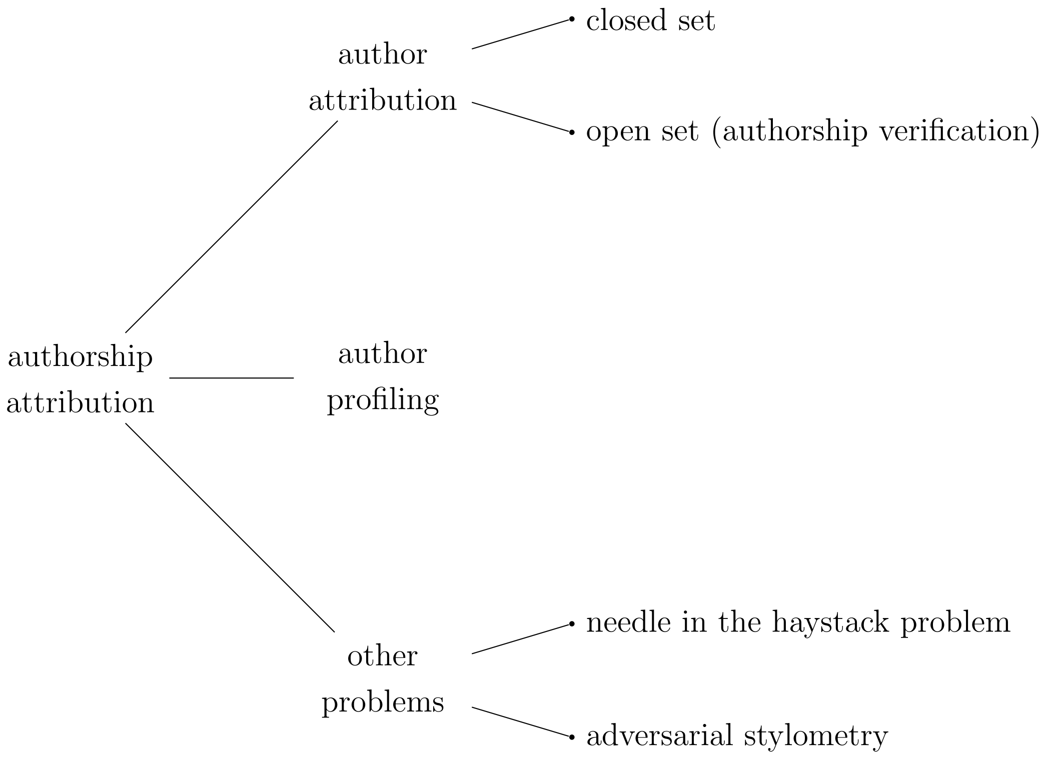


Figure 2.2: Authorship attribution types

age, gender, education level, region of language acquisition) or other information related to the author. In addition to those three problems, there is a case called *needle-in-haystack problem* that might be the hardest problem among all the authorship attribution problems. In this case, the classifier needs to identify the correct author of a text from thousands of author candidates. To make it harder, the classifier is only provided with a very small sized training data for each of the candidate authors. Later, advances in authorship attribution have raised concerns about applying attribution methods to deceptive writing and leads to the new problem space called *adversarial stylometry*. In addition to those problems, this section also discusses *cross-topic and cross-genre authorship attribution* which are still challenges in the attribution field.

2.3.1 Closed Set Authorship Attribution

Let $D = d_1, d_2, ..., d_n$ be a set of documents and $A = a_1, a_2, ..., a_m$ a fixed set of candidate authors. The task of closed set authorship attribution is to assign an author to each of the documents in D. This case is similar to text classification tasks such as sentiment analysis and topic classification where there are a predefined set of classes. Early attribution work focused on the closed set attribution task with a small number of author candidates, while more recent work tries to address the case with a larger number of authors.

One of the existing problems in the closed case attribution is the difficulty in differentiating between three factors: author, genre, and topic. In addition, it is also a challenging task to find the best stylistic features that can capture only author style information. Past research reported a wide variety of features that can be used to distinguish authors. However, usually because of a lack of systematic research, it is difficult to determine whether a high attribution performance was obtained because of the nature/genre of the text, the use of stylistic features or both of them. Authorship attribution is often biased with genre and topic identification. Function words which it is claimed capture only stylistic information, have been proved to have great potential for capturing thematic information in the document (Clement and Sharp, 2003; Mikros and Argiri, 2007). On the other hand, content words that have been shown to be useful for topic/genre classification are also reported to be effective features for distinguishing author (Koppel et al., 2009).

2.3.2 Open Set Authorship Attribution (Authorship Verification)

Given a pair of documents (X,Y), the task of author verification is to identify whether the documents have been written by same or different authors. The authorship verification task is significantly more difficult than authorship attribution. Verification does not learn about the characteristic of each author, but rather about the differences between a pair of documents. The problem is complicated by the fact that an author may consciously or unconsciously vary his/her writing style from text to text (Koppel and Schler, 2004). There has been limited research on the authorship verification problem. In general, works in authorship verification share similar approaches to those used for plagiarism detection (Stein and Meyer Zu Eissen, 2007; Stamatatos, 2009b; Zechner et al., 2009).

Author verification methods are either *intrinsic* or *extrinsic* (Stamatatos et al., 2014). The main difference between those approaches is the usage of additional documents to help the verification process. Intrinsic methods use only the provided documents (in this case known and unknown documents) to determine whether they are written by a similar author or not. Most of author verification work falls into this category. There are two common techniques used in the intrinsic method: machine learning and similarity-based approaches. Given a pair of documents (x,y), the similarity-based approach assigns the pair to a similar author if the similarity score exceeds a certain threshold. Standard vector similarity measures such as cosine similarity, Euclidean distance, min-max measure, and Manhattan distance is commonly used to generate the score. A similarity-based method can be seen as the simplest verification approach. However, it does not seem to work well, since this method tends to neglect the fact that a document's similarity is not only determined by the author's style but also by other factors such as genre and topic (Koppel and Winter, 2014).

The second intrinsic method: machine learning approaches, utilize the labeled document pairs to construct a model that can be used to classify the unlabeled pairs. Many verification works applied this approach with different learning algorithms. For example, Fréry et al. (2014) employed optimized decision trees with several representations of the texts. Their result brought them 2$^{\text{nd}}$ place in the PAN CLEF Challenge 2014 with an overall AUC-ROC[3] of 70.7% and C@1 (Peñas and Rodrigo, 2011) of 68.4%. SVM (Koppel and Winter, 2014), K-NN (Jankowska et al., 2013), Fuzzy C-means Clustering (Modaresi and Gross, 2014) are among the algorithms commonly used in verification.

In contrast to the intrinsic method, extrinsic techniques try to convert the verification problem into binary classification task by generating a large set of impostor or distractor documents which act as negative examples (Stamatatos et al., 2014; Koppel and Winter, 2014). A pair of documents will be identified as written by a similar author if the similarity score between those two is greater than the impostors (Seidman, 2013). Several verification approaches (Koppel and Winter, 2014; Seidman, 2013; Khonji and Iraqi, 2014; Mayor et al., 2014) used extrinsic technique and obtained better results than intrinsic approaches. As an example, Koppel and Winter (2014) compared the verification result of three different approaches, including the similarity based method, supervised method and impostor method. The results showed that the impostor method outperformed two other methods by obtaining an accuracy of 87.4%. However, choosing the impostor set and how many impostors to use are very critical issues that need to be dealt with. Koppel and Winter emphasized that impostor quality, impostor quantity and score threshold need to be optimized to get the proper balance of false positive and false negative in the result. They also mentioned that the impostor method still could not perform well when the genre and/or topic of the documents are different.

Since 2011, authorship verification has become one of the main tasks in the evaluation lab on uncovering plagiarism, authorship and social software misuse

[3] area under the ROC curve

(PAN CLEF Challenge)[4] (see details in Section 2.4). From time to time, the organizer tries to challenge the participants by increasing the level of difficulty of the task. For example in 2015, the participants were provided with a very limited number of known documents which had a different genre/topic to the unknown documents. However, this task represents more the real world application where variables such as genre, topic, and number of training document can not be controlled. The focus of PAN on author verification obviously gives a positive impact on this field. Despite the fact that author verification is still an unsolved task, there is significant progress on the development of new corpora, new methods and an evaluation framework in this area (Stamatatos et al., 2014).

2.3.3 Author Profiling

Previous work in author profiling mostly focused on the age, gender, and demographic profile of the authors. Some of the works also focused on the native language profiling and identification of personality types. Research in author profiling has applied various features and techniques to different types of dataset. For example, Koppel et al. (2002) used a large number of content-independent features consisting of 405 function words and a list of n-grams of part-of-speech to identify gender in 920 English documents of the BNC corpus. Optimal performance (82.6% on accuracy) was obtained when a combination of function word and part-of-speech n-grams were applied to non-fiction documents. In addition, Koppel et al. found the interesting fact that there is strong difference in usage of determiners, negation, pronoun, conjunction and preposition between male and females either in fiction or non-fiction documents. More recent work by Johannsen et al. (2015) comes to similar conclusions. Johannsen et al. conducted a study of syntactic variations among demographic groups across five different languages. Their results show men use numerals and nouns more than women, while on the other hand women use VP conjunction more frequently than men. For age, it has been found that younger groups use nouns more often, while the older age groups seem to use prepositional phrases more.

Work by Schler et al. (2006) studied the effects of age and gender on blogging. In contrast to Koppel et al., they used content-based features along with style-based features to identify an author's gender and age on 1,405,209 blog entries. Schler et al. used simple content words and LIWC's special class words (Pennebaker et al., 2001). By analyzing content-words that have high frequency and information gain, they concluded that male bloggers tend to write more about

[4] http://pan.webis.de/

politics, technology, and money while female bloggers prefer to discuss their personal life. However, both Koppel and Schler have agreed that stylometric features provide more information about the gender of the author despite the fact that there is a big difference in content between male and female.

Gender-linked features are also useful for identifying gender. Work by Cheng et al. (2011) applied nine gender-linked cues including use of affective adjectives, exclamation, expletives, hedges, intensive adverbs, judgmental adjectives and uncertainty verbs. Past studies (Jaffe et al., 1995; Mulac et al., 1990) showed that different genders of authors have their preferences on using gender-linked cues. For example, compared to female, male authors rarely use emotionally intensive adverbs and affective adjectives in their writing. Men often use first-person singular pronouns to express independence and assertions.

In PAN-2014, there were 10 submissions for the task of author profiling (Rangel et al., 2014). Among the submissions, López-Monroy et al. (2014) successfully obtained the overall best performance on average for English and Spanish tasks. In their proposed approach, they built term and document vectors which represent the relationship of the term/document with the author's profiles (e.g. male, female). In addition, they generated author's subprofiles (e.g. young-gamer females, housewife females) and re-computed the term/document representations using subprofiles as the new target profiles. Their results outperformed the vast majority of the teams who applied common author profiling features such as stylistic and content-based features.

2.3.4 *Needle-in-a-haystack* Attribution Problem

Studies in the authorship attribution field mostly focus on small set problems where there is a small number of authors and a large amount of training data. In most cases, the problem can be addressed using supervised classifiers and stylometric features as the input. However, more recent work found those standard techniques performed less reliably in terms of accuracy and computation time when applied to real-world applications like forensic investigation. The challenges faced in the *needle-in-a-haystack* problem is much closer to the practical applications. In this problem, the goal is to identify the author of a document but provided with thousands of potential candidates and a small training size.

A systematic study of the effect of author set size and training data size was conducted by Luyckx and Daelemans (2010). The three evaluation data sets were used in their experiments consisted of an English data set with 13 authors, a Dutch data set with eight authors and a larger Dutch data set with

145 authors. In order to see the effect of author set and data set size on attribution performance, they gradually increased the number of authors and training data in separate experiments. In line with the results of Koppel et al. (2011), the experiment showed a significant decrease in the attribution accuracy when there are more candidate authors. This trend occurred in each of the datasets regardless of their language, number of topics or the training size. In addition, they found an important aspect of the feature set: similar types of features e.g. character n-grams tend to perform well on both small and large numbers of authors.

Instead of using standard attribution approaches, Koppel et al. (2009) addressed the *needle-in-a-haystack* problems by applying information retrieval methods. First, they constructed the dataset by taking snippets of at least 500 words from 20,000 blogs. 10,000 blogs were used in the training and the remaining as the test set. Each snippet was then represented in three different TfIdf representations based on 1,000 most common words and another based on style features. The author of an unlabeled text is predicted by taking a candidate whose known work has the highest similarity with a given snippet. The three content representations produced performance between 52% and 56%, while the style representation performed much worse by producing only 6% in accuracy.

The scalability issue certainly will be a great challenge for people in the authorship attribution field. However, Luyckx and Daelemans (2010) emphasized that scalability is only one of a number issues faced in large-scale authorship attribution. This issue should expand our understanding that research in authorship attribution should be done by considering factors in a real world situation. Another example of the challenge faced in large-scale authorship attribution is the development of an adaptive attribution method (Stamatatos, 2009a). In the practical application, it is difficult to have an ideal version of corpora with a small number of candidate authors, big size of training data and a controlled genre/topic for training and testing data. Thus, an adaptive and robust attribution method is certainly needed. An attribution method is needed that is not only robust to the scalability issues but also can be trained on training data from one topic/genre and tested on different topic/genre dataset.

2.3.5 Adversarial Stylometry

Most previous studies in authorship attribution assumed that authors write in their original writing style without any intention of modifying it. It has been shown that current authorship attribution methods could achieve relatively high accuracy in identifying authors (Abbasi and Chen, 2008). Advances in authorship attribution have raised concerns about applying attribution methods to adversarial writing and this has led to the new problem area called *adversarial stylometry*. According to Brennan et al. (2012), adversarial writing can be in the form of *obfuscated writing*, where the subject tries to modify his original writing style; *imitated writing*, where the subject is copying another person's writing style; and *translated writing*, where machine translation is used to modify the original text. The main goal of adversarial writing is to hide the true author's writing style.

Adversarial writing could be useful as it resolves the privacy and security issue on the internet. As claimed by Rao and Rohatgi (2000), privacy and security approaches focus on anonymizing proxies, cryptographic, and traffic shaping techniques, yet tend to ignore the source of information itself. In their works, Rao and Rohatgi demonstrated how stylometry provides solutions to the privacy problem. Adversarial writing may lead to fraud when people obfuscate their writing style to hide their true identity.

Adversarial stylometry can be considered as a new task in the authorship attribution field. There have been few studies in adversarial stylometry and those that have appeared mainly studied the impact of the adversarial attack on authorship attribution performance. As an example, Kacmarcik and Gamon (2006) observed the effect of changing feature vector values on attribution accuracy. In their experiment, they only used the most frequent and highly ranked word features. Using an SVM classifier, they found that not much effort was needed to obfuscate a text document, as only an average of 14 changes in the feature vector were required per 1000 words to give the mis-attribution effect.

It has been found that current attribution methods seem not resilient to the adversarial attack as reported by Brennan and Greenstadt (2009). In their experiment, three different attribution methods including chi-square, neural network, and a synonym-based classifier with a different set of features for each of them were applied to identify author on obfuscated and imitated documents. A corpus was created from the writing of 15 individual authors. Each of the authors submitted three types of documents: original, obfuscated and imitated text. The results showed that there was significant decrement on the accuracy of each attribution method when dealing with the adversarial attack. The drop in accuracy increased as the number of candidate authors increased. Similar results were shown in other adversarial stylometry works (Juola and Vescovi, 2010; Brennan et al., 2012; Afroz et al., 2012).

Attribution methods might be very vulnerable to adversarial attack; however detecting deceptive writing is not a hard task. Using a large feature set, Afroz et al. (2012) successfully achieved 96.6% in accuracy for distinguishing deceptive document from the regular one. In addition, they found that detecting an obfuscation attack was harder than detecting an imitation attack. Interestingly, they reported that in deceptive writing, a person tends to use simpler words with fewer syllables, shorter and less complex sentence.

There is relatively little work that has been done in the adversarial stylometry area. Among of them is the work by Brennan (2012) who developed two adversarial corpora: Brennan-Greenstadt and Extended Brennan-Greenstadt adversarial corpus. The first corpus was created based on a survey conducted by Drexel University. This corpus contains three basic elements: the first element is a preexisting sample of writing from 13 nonprofessional writers, where each of them submitted at least 6500 words. In order to eliminate slang and abbreviations, each writing sample had to be formal writing such as an essay for school, report for work or other academic and professional correspondences.

The second element is a 500 word obfuscation passage on a specific topic, while the last element contains the sample of imitation passages from the participants where they tried to imitate another author's style (in this case Cormac McCarthy). On the second and third elements, only 12 authors participated. In order to conduct more robust analysis, Brennan and Greenstadt created a larger and more diverse adversarial corpus called an Extended Brennan-Greenstadt adversarial corpus. On this corpus development, submission quality is the main concern. Each submission had to strictly follow the directions given. Among 100 submissions, only submissions from 45 individual authors were taken to construct this corpus.

2.3.6 Cross-Topic and Cross-Genre Authorship Attribution

Cross-topic and cross-genre authorship attribution is another challenge in the attribution field. In this case, the genre and/or topic may differ significantly between the training and test documents. This task is more realistic since in real world applications the genre/topic of the documents can not be controlled. Kestemont et al. (2012) applied unmasking methods (Koppel et al., 2007) to authorship verification across genres (prose and theater play). Given two documents A and B, unmasking method works by generating a curve which demonstrates the accuracy degradation when k-most useful features are removed. A sudden and dramatic degradation curve indicates those documents were written by different authors. In contrast, if the degradation curve is slow and smooth, document A and B were potentially written by the same author.

Compared to the result of intra-genre authorship verification, there is significant degradation of performance for cross-genre verification. Work by Stamatatos (2013) studied the effectiveness of character n-gram and word features for cross-topic and cross-genre authorship attribution. First, cross-topic attribution was examined where a political text is used as training data and various thematic texts (society, world, and U.K.) as the testing set. Note that, the training and testing set were still in the same genre. The same scenario was applied for cross-genre attribution. The classifier was trained on politics texts and tested on book reviews. Both cross-topic and cross-genre attribution showed similar results. The performance decreased considerably compared to intra-topic/intra-genre attribution. Stamatatos observed that in cross-topic/cross-genre attribution, low-frequency features should be avoided as they reduced the effectiveness of the attribution models.

Sapkota et al. (2014) used multiple cross-topic documents to train the cross-topic authorship attribution model. From their results, it was found that their proposed models could significantly improve the performance of cross-topic authorship attribution, which is also an indication that authors maintain a consistent writing style regardless of their topic preference. In addition to that, they presented an analysis of feature sensitivities towards the change of topics. By comparing four different attribution features, they concluded that character n-grams have higher discriminative power in this task.

2.4 Evaluation Forum

According to Rudman (2012), there are still major shortcomings in authorship attribution research that need to be addressed. One criticism is the lack of continuity as researchers do not seem to have any long-range commitment to work on one problem (Rudman, 1997). He also noted that only a small percentage of attribution works can be reproduced. The lack of consensus on the standard attribution datasets is also another problem to be addressed. The PAN Evaluation Lab tries to offer the solution for the above problems by organizing annual authorship attribution shared tasks. PAN provided standard datasets and mechanisms to evaluate each of the proposed attribution approaches which make it easier to do the benchmarking on the results.

Since 2011, the author identification task has been part of PAN (Plagiarism, Authorship, and social software misuse) evaluation which is hosted by the CLEF initiative (Conference and Labs of the Evaluation Forum). In each year, PAN covers different attribution tasks; starting with standard authorship attribution to more advanced tasks such as author clustering and obfuscation. The challenge has been developed to become more similar to realistic applications. For example, PAN 2016 focused on author diarization (Rosso et al., 2016). This task is to identify different authors within a single collaborative work (e.g., paper with many authors) or the result of plagiarism. In 2017, a similar task was introduced called style breach detection (Tschuggnall et al., 2017). The main goal of this task is to identify the exact position in a collaborative document where the authorship changes. Unlike author diarization where the training data is available, the style breach detection task provides no training data. In addition, no information can be gained from web search. The style breach detection task can be considered as a text segmentation problem with the focus on detecting switches of writing style, disregarding the specific content or topic.

An author masking/obfuscation task was also introduced for the first time in 2016. The main goal of this task is to check the robustness of current state-of-the-art attribution methods against the obfuscation techniques. As illustrated in Figure 2.3, author masking is an opposite task of author verification (see Section 2.3.2). Given two documents written by the same author, author masking works by paraphrasing one of the documents so that the author can not be identified anymore. In contrast to that, the author verification task has to verify whether two documents has the same author. Thus, the development of new approach in author obfuscation will influence the capabilities of authorship verification.

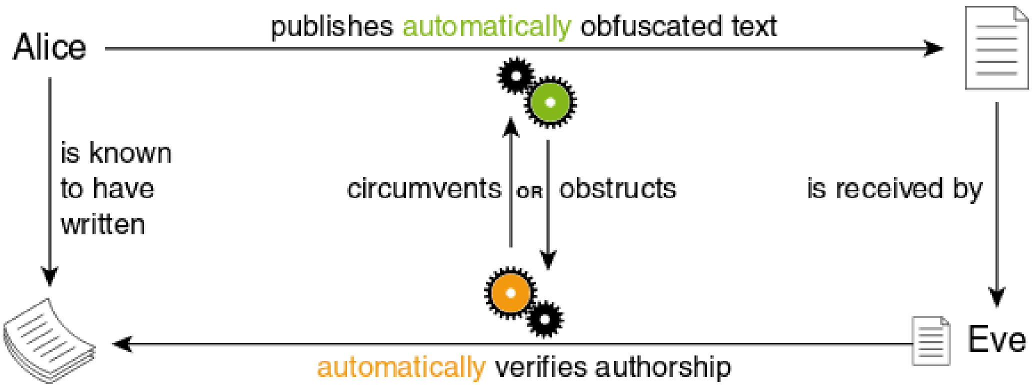

Alice
publishes automatically obfuscated text
is known
to have
written
circumvents
OR
obstructs
is received by
Eve
automatically verifies authorship

Figure 2.3: Schema of author masking/obfuscation and verification (Potthast et al., 2016b)

In addition to the diversity of the tasks, PAN also encourages the participants to develop more adaptive attribution methods by providing datasets in various natural languages (Dutch, English, Greek, and Spanish) and genres/topics (essays, reviews, novels, opinion articles, e-mails). Details of PAN's authorship attribution tasks are provided in Table 2.2.

Year	Tasks	Languages	Topic/genre
2011	Author attribution, author verification	English	Email
2012	Author attribution, author verification, author clustering	English	Fiction book collections
2013	Author verification, author profiling	English, Greek, Spanish	Computer textbooks, news articles, newspaper editorials, short fictions
2014	Author verification, author profiling	Dutch, English, Greek, Spanish	Essays, reviews, novels, opinion articles
2015	Cross-topic and cross-genre author verification, author profiling	Dutch, English, Greek, Spanish	Essays, reviews, opinion articles, play script
2016	Author clustering, author diarization, author profiling, author obfuscation	English, Dutch, Greek	newspaper articles, text from online forum
2017	Author clustering, style breach detection, author profiling, author obfuscation	English, Dutch, Greek	Newspaper articles, reviews

Table 2.2: Details of the PAN's authorship attribution shared tasks

However, regardless of their attempts to create diversity in the task, we observed that the released datasets are rather small in size. As an example, in the PAN 2015 authorship verification task, the organizer provided only 100 training cases per language. We argue that the small size data limits the participants from exploring more advanced approaches which usually require large amounts of training data. The techniques proposed by the participants for each of PAN's shared tasks illustrates how authorship attribution methods develop over time. However, for most cases, they are dominated by the combination of supervised machine learning algorithms such as SVM with stylometric features. Some of the participants chose to focus more on feature engineering in order to identify the most effective features (Hürlimann et al., 2015). There are also neural network-based methods that have been proposed. For example in 2015, Bagnall (2015) proposed a character-based language model using Recurrent Neural Network (RNN) for author verification tasks. He argued that the model is suitable for small amounts of data and can capture idiosyncratic usage in the text. The proposed approaches obtained the best-performing results among all the participants. However the author also reported that the computational cost was high.

2.5 Summary

This chapter presented an overview of authorship attribution. It started with a review of the authorship attribution task. and showed how the authorship attribution techniques have developed over time. Then, different authorship attribution features were compared. One of the important findings from the previous work is that features which are usually discarded in the content-based classification task can be useful for the style-based task. The usage of those features in the text which could not be controlled consciously by the author make it as a useful clue to capture the author's writing style.

The chapter proceeded with a review on several authorship attribution forms; including closed set authorship attribution, authorship verification, author profiling, adversarial stylometry, the *needle-in-the-haystack* problem and cross-topic or cross-genre authorship attribution. The last section provided reviews of the PAN authorship attribution evaluation forum.

Chapter 3

Exploring the Most Useful Features for Authorship Attribution

Authorship attribution has been extensively studied and a wide range of features explored (see Section 2.2) (Stamatatos, 2013; Schwartz et al., 2013; Seroussi et al., 2013; Hürlimann et al., 2015). However, there has been a lack of analysis of the behavior of features across multiple datasets or when using a range of classifiers. Consequently, it is difficult to determine which types of information will be most useful for a particular authorship attribution dataset. There have been some attempts in feature exploration, e.g. (Guthrie, 2008; Stamatatos, 2009a; Brennan et al., 2012; Sapkota et al., 2015). Guthrie (2008) and Brennan et al. (2012) examined several types of stylistic and linguistic features used to characterize writing. Sapkota et al. (2015) attempted to evaluate the function of different character n-gram subgroups for authorship attribution. However, the work mostly focused on the overall effectiveness of features without considering the characteristics of the datasets to which they were applied.

Authorship attribution is a unique task which is closely related to both the representation of individuals' writing styles and text categorization. In some cases, where there is a clear topical distinction between the documents written by different authors, content-related features such as those used in text categorization may be effective. However, style-based features are more likely to be effective for datasets containing a more homogeneous set of topics. Many previous studies (Peng et al., 2003; Koppel et al., 2011; Sapkota et al., 2015; Schwartz et al., 2013; Sari et al., 2017; Shrestha et al., 2017) have concluded that, among the large number of features that have been applied to the authorship attribution problem, using character n-grams features often produces good accuracy. Thus, character n-grams have become the *go-to* features for this task to capture both an author's topical preferences and writing style.

This chapter explores how the characteristics of an authorship attribution dataset affect the usefulness of different types of features. We carry out an analysis of four datasets that have been previously used for this task: Judgment, CCAT10, CCAT50 and IMDb62. Three types of features are considered in this study: *style*, *content* and *hybrid* (a mixture of the previous two types). The analysis indicates that features intended to capture topical preferences are most useful when there is a clear topical distinction between authors and that this can be predicted by analysing the output from a topic model. In contrast to previous work, this study finds that character n-grams do not perform equally well in all datasets. The analysis holds for authorship attribution models using discrete and continuous representations. Using topic modeling and feature analysis, the most effective features can be successfully predicted for three of the four datasets.

The remainder of this chapter is structured as follows. We start by describing details of the datasets used in this thesis (Section 3.1). In Section 3.2, the dataset analysis using topic modeling is presented. In this analysis, we utilize Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to model topical interest between authors. We then present the feature analysis in Section 3.3. Feature ablation studies are conducted using a total of 728 features including lexical, syntactic, character and content-based features. The ablation studies are performed for both continuous and discrete representation. Then, we present the existing work in Section 3.4, followed by a summary in Section 3.5.

3.1 Datasets

In this section, we present an overview of datasets involved in this thesis. Instead of creating our own data, we choose to perform experiments with the available corpora which have been commonly used in previous literature. By conducting experiments on the available datasets, we can benchmark our models against the previous results. Furthermore, it ensures evaluation comparability between approaches which are useful to form a more solid foundation for the authorship attribution field. Another major advantage of using commonly used datasets is the warranty of the datasets' quality. Rudman (2012) pointed out that the corpus used in authorship attribution needs to be constructed carefully by taking attention to various points. As an example, texts for a corpus need to be the authentic writing of an author. Thus, texts that are obtained from the web

	Judgment	CCAT10	CCAT50	IMDb62
genre	legal judgments	newswire		movie reviews
# authors	3	10	50	62
# total documents	1,342	1,000	5,000	79,550
avg characters per document	11,957	3,089	3,058	1,401
avg words per document	2367	580	584	288

Table 3.1: Dataset statistics

or internet need to go through specific pre-processing steps such as removing all extraneous text (i.e. additional text from editor or commentator), applying encoding, regularizing or lemmatizing. This thesis utilizes four main datasets that have been commonly used in previous attribution work: Judgment, CCAT10, CCAT50, and IMDb62. These datasets represent a range of characteristics in terms of the number of authors, topic/genre and document length (see details in Table 3.1).

Judgment (Seroussi et al., 2011). The Judgment dataset was collected from legal judgments of three Australian High Court judges: Dixon, McTiernan, and Rich. This dataset was created to verify rumors of Dixon's ghost-writing attributed to McTiernan and Rich. Judgment is an example of a traditional authorship attribution dataset where there are only a small number of authors with relatively long text in a formal language (Seroussi et al., 2013). Due to the genre of the datasets, Seroussi argued that attribution approaches that consider author's writing style are more likely to obtain better performances rather than methods that rely solely on content-based features. The dataset comes in pre-processed form where all dates and quotes were removed to ensure that only the actual authors' language is left. In this thesis, we follow Seroussi et al. (2013) by using only undisputed judgments which were indicated by the periods when only one of the three judges served on the High Court (Dixon's 1929–1964, McTiernan's 1965–1975, and Rich's 1913–1928 judgments). Judgment has an imbalanced number of documents per author with 902 docs from Dixon, 253 docs from McTiernan and 187 docs from Rich. As the dataset does not come with separate train-test partitions, we follow the previous work by using 10-fold cross-validation in our experiments. Figure 3.1 shows a snippet from the Judgment dataset.

> the plaintiff , who is the mortgagee of land under the transfer of land act vict. , claims in this action that she , instead of the mortgagor , is entitled by virtue of her rights as mortgagee to receive from the commonwealth the compensation which it agreed with the mortgagor pursuant to reg . d of the national security (general) regulations to pay him . the facts alleged to support this claim are briefly that the minister validly took possession of one parcel of thirty acres and another of sixty-two acres of the mortgaged land pursuant to reg . of those regulations . the commonwealth agreed with the mortgagor to pay him " " at the rate of four pounds per week " " and fifty-five pounds per annum , payable monthly , " " the compensation was " " of these two parcels of the mortgaged land ...

Figure 3.1: A snippet from Judgment dataset

CCAT10 (Stamatatos, 2008). This dataset is a subset of Reuters Corpus Volume 1 (RCV1) (Rose et al., 2002) and consists of newswire stories by ten authors labeled with the code CCAT (which indicates corporate/industrial news). The corpus was divided into 50 training and 50 test texts per author. In our experiments, we follow prior work (Stamatatos, 2013) and measure accuracy using the train/test partition provided.

CCAT50 This corpus is the larger version of CCAT10. In total, there are 5,000 documents from 50 authors. As for CCAT10, for each of the author there are 50 training and 50 test documents. A sample text from CCAT10/CCAT50 datasets is shown in Figure 3.2.

> Several countries, mostly Asian, are resisting what they see as a U.S.-driven push to get labour rights on to the WTO agenda and to use the organisation's disputes court to erode the developing world's low-cost labour edge. When asked if Malaysia would permit a WTO study on labour rights, International Trade and Industry Minister Rafidah Aziz told reporters:"No, no, no way. There is no place for labour issues at the WTO".

Figure 3.2: A snippet from CCAT10/CCAT50 datasets

IMDb62 (Seroussi et al., 2010). IMDb62 dataset consists of 62,000 movie reviews and 17,550 message board posts from 62 prolific users of the Internet Movie database (IMDb, `www.imdb.com`). Each user wrote 1,000 movie reviews and a different number of message board posts, the topics of which may also be about movie, television, music, and other topics. Among the datasets that were used in our experiments, IMDb62 has the largest number of authors and documents.

This dataset allows us to examine our approach in medium-scale authorship attribution with similar themes of texts and plentiful training data. Similar to Judgment, 10-fold cross validation is implemented in the experiments. Compared to Judgment and CCAT's, the language used in IMDb62 is considered less formal (see Figure 3.3)

> Poor film dealing with a brother's scheme to get out of jail. He changes places with his brother who looks like him.Jack Palance as always is intriguing. However, the problem here is that he has such poor written material to work with.Harold J. Stone comes off as a heavy as a prison guard who has larceny in his heart but has the tables turned on him.We see Palance as a sympathetic brother who helped the latter through college...

Figure 3.3: A snippet from IMDb62 dataset

3.2 Dataset Analysis

In this section, analysis of the data sets using topic modeling is presented. The aim of this analysis is to quantify topical divergences between authors in each of the datasets. The motivation for this is that certain datasets may have clear topical preferences between authors which cause authorship attribution to be biased towards topic classification. Therefore, topic modeling can help assess the topical dis-similarity between authors.

3.2.1 Analysis using Topic Modeling

We perform topic modeling using Latent Dirichlet Allocation (LDA) (Blei et al., 2003). LDA is a generative probabilistic model of a corpus. Documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. Figure 3.4 shows the graphical representation of LDA using a plate diagram.

The graph can be explained as follows, given:

- $D = \mathbf{x}^1 ... \mathbf{x}^M$ is dataset containing M documents.
- $\mathbf{x} = [x_1 ... x_N]$ document with **N** words.
- θ^m is the topic proportion for the m-th document.

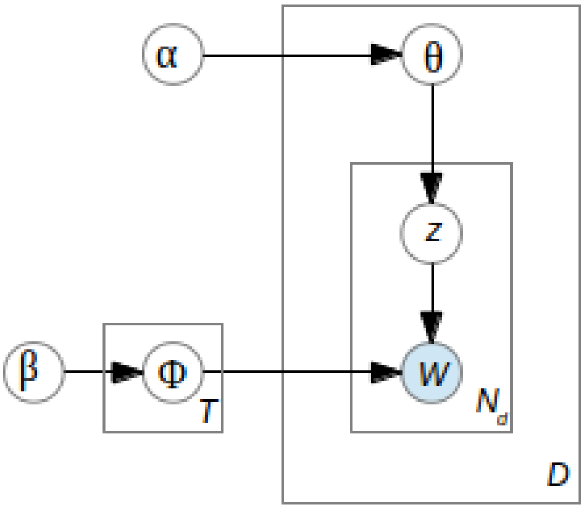


Figure 3.4: Graphical model representation of LDA

- $\Phi_{1:T}$ are the topics, where each Φ_t is a distribution over words in the vocabulary.
- $\mathbf{z}^m$ is the topic assignment for the m-th document.
- α is the parameter of the Dirichlet prior on the per-document topic distribution.
- β is the parameter of the Dirichlet prior on the per-topic word distribution.

Then the topic probability of document $\mathbf{x}^m$ given the hyper-parameters α and β is given by the following equation:

$$P(\mathbf{x}^m, \mathbf{z}^m, \theta^m, \Phi_{1\ldots T}|\alpha, \beta) = Dir(\theta^m|\alpha)\prod_{t=1}^{T} Dir(\Phi_t|\beta)\times\prod_{n=1}^{N} Mult(z_n^m|\theta^m)Mult(x_n^m|\Phi_{z_n^m}) \tag{3.1}$$

Parameters θ and Φ are estimated using Gibbs Sampling algorithm (Griffiths and Steyvers, 2004).

Assuming a trained topic model over an authorship attribution dataset D, if C_a is the set of documents written by author a and σ_i is the topic distribution for the i-th document in C_a, then we estimate the topic distribution for a particular author as follows:

$$\theta_a = \frac{\sum_{i=1}^{|C_a|} \sigma_i}{|C_a|} \tag{3.2}$$

Following this, the difference between two author's topic probability distributions is calculated using the Jensen-Shannon Divergence (JSD) (Cover and Thomas, 2006):

$$sim(P, Q) = JSD(P||Q) \tag{3.3}$$

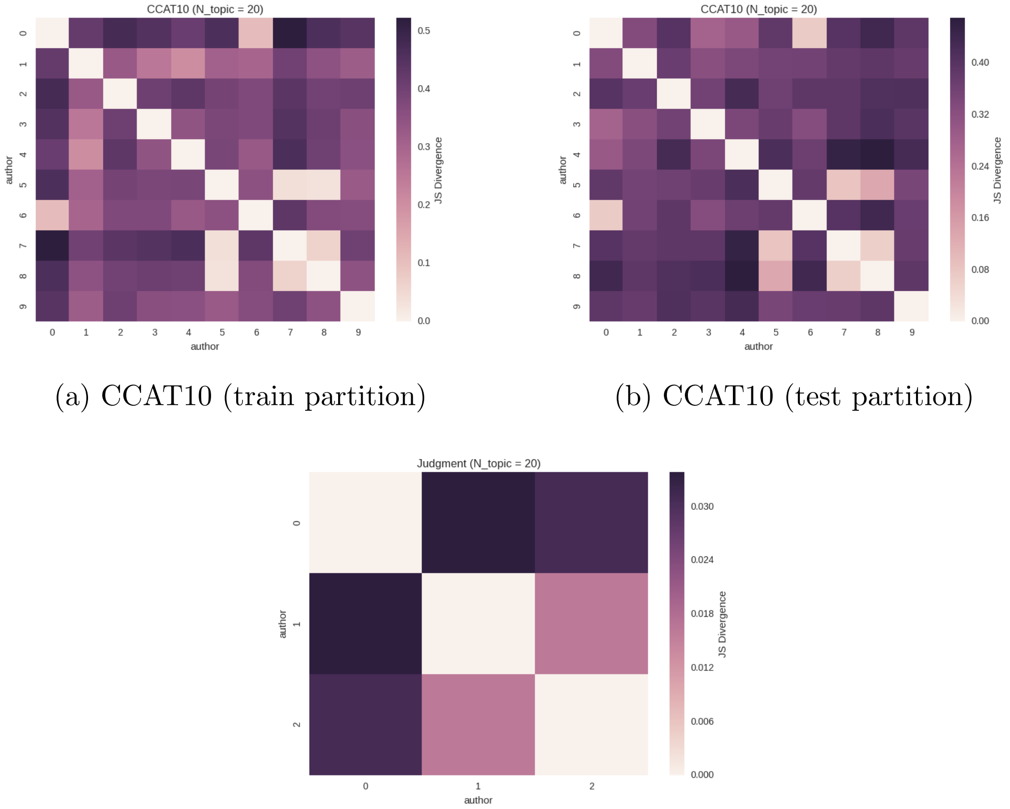
3.2.2 Results and Discussion

Table 3.2 shows the average of JSD for all author pairs in each of the datasets having trained a topic model with a different number of topics. High JSD scores indicate more topical diversity between authors in the dataset. The CCAT datasets, which contain on-line news, have higher scores compared to Judgment and IMDb62. The scores for CCAT50 and CCAT10 are similar, despite the fact that the first dataset contains five times the number of authors of the second. The consistency of this comparison across different numbers of topics indicates that this method of assessing content similarity between authors is robust with respect to tuning this parameter. Judgment has the lowest score across the four datasets indicating that the authors discuss the most similar topics. Finally, scores for the IMDb62 dataset obtained were slightly higher than those for Judgment. Differences in scores for IMDb62 are due to the authors' preferences, some commented on the story while other commented on the characters of the movie. Furthermore, from the results we observe that the genre of the datasets influences the topical divergences between authors. Datasets constructed from on-line news tend to have higher topical diversity. In contrast to that, legal judgments and movie reviews have limited topic variances.

Confusion matrices were created to further analyse differences between authors. These matrices were generated after running LDA with 20 topics for 1000 iterations. Similar patterns was observed using different numbers of topics. For CCAT10 and CCAT50, separate matrices were generated for both train and test partitions. Darker color indicates higher JSD between two authors. While

n_topic	Judgment	CCAT10 (tr)	CCAT10 (ts)	CCAT50 (tr)	CCAT50 (ts)	IMDb62
3	0.0056	0.1889	0.1936	0.1526	0.1728	0.1000
10	0.0148	0.3030	0.2872	0.2618	0.2627	0.1471
20	0.0180	0.3224	0.3214	0.3067	0.2956	0.1617
30	0.0256	0.3485	0.3319	0.3151	0.3158	0.1627
40	0.0272	0.3485	0.3360	0.3293	0.3262	0.1681
50	0.0281	0.3527	0.3459	0.3369	0.3345	0.1634

Table 3.2: Average JS Divergence for each number of topics
(tr: training data; ts: test data)



(a) CCAT10 (train partition)

(b) CCAT10 (test partition)

(c) Judgment

Figure 3.5: Author topic distribution (n_topic = 20)

lighter color indicates higher topical similarity between authors. For example in CCAT10's train set (Figure 3.5a), authors 5, 7, and 8 shared similar topical interests related to *China* and *Beijing*, but from different points of view. The majority of documents written by authors 7 and 8 discuss regional events within

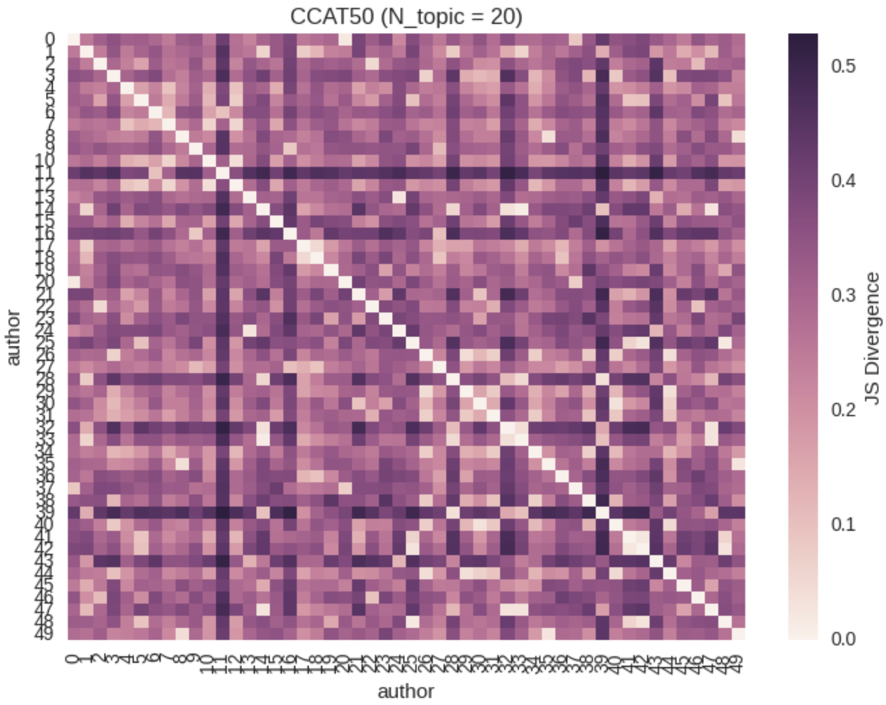


Figure 3.6: Author topic distribution (n_topic = 20) in CCAT50 (train partition)

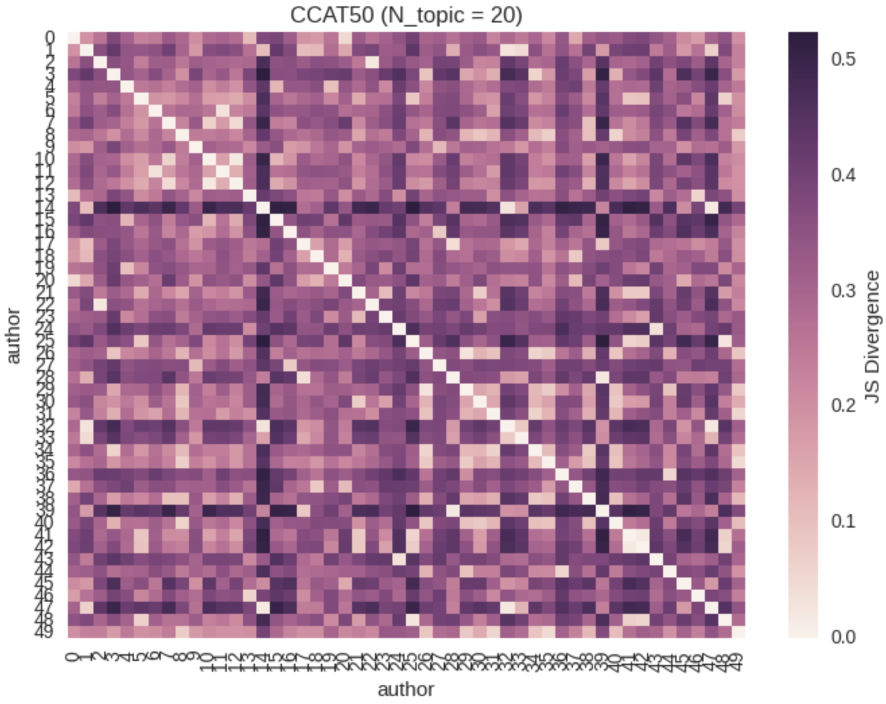


Figure 3.7: Author topic distribution (n_topic = 20) in CCAT50 (test partition)

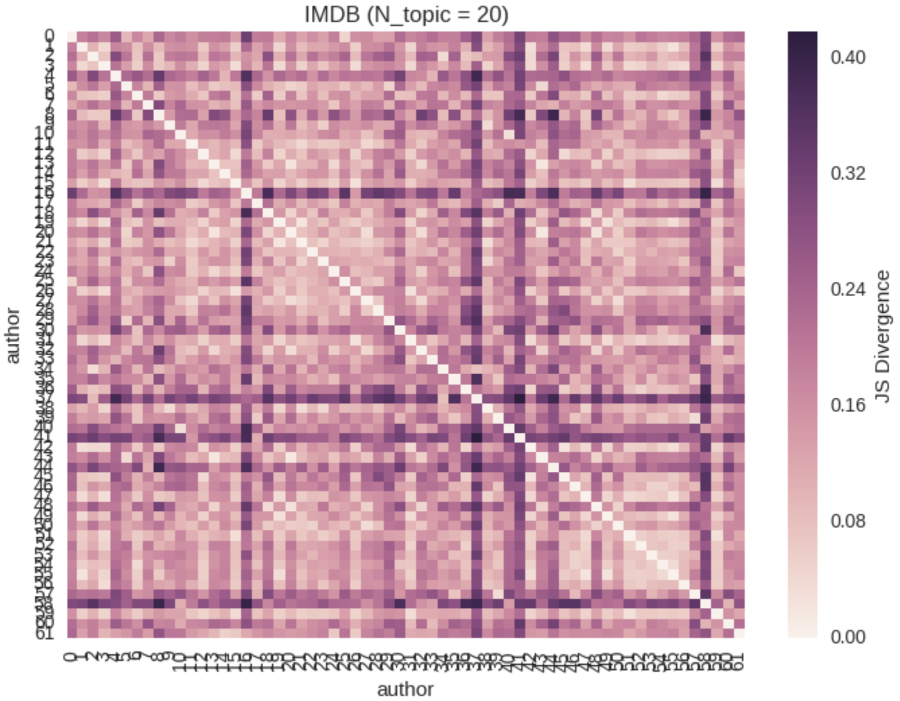


Figure 3.8: Author topic distribution (n_topic = 20) in IMDb62

China (Figure 3.9, Topic 9). On the other hand, author 5 reported more about international related events (Figure 3.9, Topic 3).

Topic 3

china beijing taiwan chinese foreign visit washington relations ties sino trade president united states war island talks nuclear links taiwanese

Topic 9

china state party beijing communist chinese official people officials government economic years jiang xinhua newspaper yuan deng li corruption law

Figure 3.9: Sample of topics in CCAT10 datasets. Topics are represented by top-20 most probable words.

In the CCAT50 dataset (Figure 3.6), one author (number 11) has very different topic preferences to the others. Articles written by author 11 mainly discuss topics related to *gold, exploration, Canada, Indonesia* which are rarely picked by the other authors. A similar pattern is found in IMDb62, as shown in Figure 3.8. Reviews by author 16 are dominated by positive comments about movies unlike other authors who tended to write negative reviews or discuss the *story* and/or *characters*. Unlike the three other datasets, authors in Judgment wrote on relatively similar topics (Figure 3.5c). We observed that there is no particular topic dominated in author's writings.

We performed another experiment to examine the influence of the number of authors to the topical divergence. The experiment was conducted on the CCAT50 and IMDb62 data sets with a number of topics equal to 20. For each number of authors, LDA was run ten times with random combinations of authors. The results are presented in Table 3.3. As can be observed from the table, in CCAT50 which has high topical divergence, adding more authors does not significantly affect the JS divergence score. A similar pattern is shown in IMDb62. The experiment with more authors causes only a small decrease in the JSD score. The results imply that a large number of authors does not guarantee that the topics will be more diverse. We argue that factors such as the genre of the dataset affects the JSD score. Although the number of authors in IMDb62 is large, they all wrote on similar topics related to movies. In contrast, CCAT10 which is constructed from newswire has higher JSD scores even though the number of authors is small. In the next section, analysis of the features is presented. Ablation studies are performed to examine whether specific features types produce better performance for a particular dataset.

n_author	CCAT50	IMDb62
10	0.2953	0.1474
20	0.2977	0.1375
30	0.2901	0.1282
40	0.2993	0.1274
50	0.2979	0.1269

Table 3.3: Average JS Divergence for different number of author

3.3 Feature Analysis

The choice of features is an important decision in the development of supervised authorship attribution methods. Previous studies proposed different sets of authorship attribution features. As an example, Stamatatos (2009a) introduced five different types of features: lexical, character, syntactic, semantic and application-specific features. Guthrie (2008) used 166 features consisting of typical stylistic features and several other features to capture emotional tone. Among those features, he found 15 features to be the most useful for authorship attribu-

Type	Group	Category	#	Description
Style	Lexical	Word-level	2	Average word length, number of short words
		Char-level	2	Percentage of digits, percentage of uppercase letters
		Letters	26	Letter frequency
		Digits	10	Digit frequency
		Vocabulary richness	2	Richness (hapax-legomena and dis-legomena)
	Syntactic	Function words	174	Frequency of function words
		Punctuation	12	Occurrence of punctuation
Content	Word n-gram	Words unigrams	100	Frequency of 100 most common word unigrams
		Words bigrams	100	Frequency of 100 most common word bigrams
		Word trigrams	100	Frequency of 100 most common word trigrams
Hybrid	Char n-gram	Char bigrams	100	Frequency of 100 most common character bigrams
		Char trigrams	100	Frequency of 100 most common character trigrams

Table 3.4: Authorship attribution feature sets

tion. Brennan et al. (2012) attempted to carry out a feature exploration. They used a simplification of Writeprints features (Abbasi and Chen, 2008) which consisted of a group of lexical and syntactic features. For a comprehensive review of authorship attribution features, see Section 2.2.

Determining the most useful features can be challenging. One way to find out the useful features is by implementing a comprehensive range of features and performing an ablation study. For this purpose, we adopted features groups from two previous studies by Stamatatos (2009a) and Abbasi and Chen (2008). These features have been widely used and proved to be effective in many attribution works. Similar to Stamatatos who introduced five features groups, Abassi and Chen proposed an extensive list of authorship attribution features called Writeprints which consist of 327 lexical, syntactic, structural and content-specific features.

We divide the features used in our experiment into three types (see Table 3.4):

- **style:** style-based features capture the writing style of the authors such as the usage of function words, digits and punctuation. We used pre-defined function words and punctuation marks (see A.1).

- **content:** content-based features consist of bag of word n-grams which are useful to map the author's topical preferences. All function words are removed when extracting these features.
- **hybrid:** in this type we use character n-grams which have been proved effective for capturing both writing style and topical preferences (Koppel et al., 2011; Sapkota et al., 2015).

Both character and word n-grams are limited to tri-grams. As the purpose of these ablation experiments is not to outperform the previous work, we used only 100 most common features for each of n-gram features. In addition, features which are produced using NLP analysis tools, such as part-of-speech and semantic features are not utilized in these experiments to avoid performance bias.

3.3.1 Feature Ablation Experiment

Ablation experiments were performed using a leave-one-out scenario. First, we conducted an experiment with all features. Then, one feature group was removed to show the effect of leaving it out. We performed feature ablation experiments using two models:

- **Feed-Forward Neural Network (FNN)**
  A single hidden layer feed-forward neural network model (FNN) was implemented. The FNN hyper-parameters including learning rates, hidden size and dropout rates were tuned on the development set for each of the datasets. Table 3.5 presents the optimal hyper-parameters for each configuration of features.

  For Judgment, CCAT10 and CCAT50, we set the number of epochs to 250 and 100 for IMDb62. For all datasets, early stopping was used on the development sets and models optimized with the Adam update rule (Kingma and Ba, 2014). Since none of the datasets have a standard development set, we randomly selected 10% of the training data for this purpose.

- **Logistic Regression (LR)**
  For Logistic Regression, the default hyper-parameter configurations from Scikit-learn (Pedregosa et al., 2011) were used.

Dataset	all features	(−) style	(−) content	(−) hybrid
Judgment	5x10^{-4};250;0.4	5x10^{-4};300;0.4	1x10^{-3};300;0.5	5x10^{-4};300;0.2
CCAT10	5x10^{-4};500;0.15	1x10^{-3};500;0.3	1x10^{-3};500;0.5	1x10^{-3};500;0.4
CCAT50	5x10^{-4};500;0.4	5x10^{-4};600;0.6	5x10^{-4};5000.6	5x10^{-4};5000.6
IMDb62	1x10^{-2};120;0.1	1x10^{-2};120;0.1	1x10^{-2};120;0.1	1x10^{-2};120;0.1

Table 3.5: Optimal hyper-parameters for each dataset with different feature configurations (from left to right: learning rate; hidden size; dropout rate)

.

Accuracy was used as the evaluation metric to measure the authorship attribution performance.

3.3.2 Results and Analysis

The results are presented in Table 3.6. The (−) symbol indicates that the respective feature type is excluded. Using the all features set, both FNN and LR produced similar accuracy. Among other feature types, removing style-based features caused the biggest drop in Judgment and IMDb62. Meanwhile, in both CCAT datasets, there was a significant decrease in accuracy when content-based features were removed. The results confirm our topic model-based analysis. Style-based features are more effective for datasets in which authors discuss similar topics, e.g. Judgment and IMDb62. As expected, content-based features are generally more effective when there is more diversity between the topics discussed in the dataset, e.g. CCAT10 and CCAT50, but are of limited usefulness when the topics are similar (particularly for the Judgment dataset). The hybrid features appear to behave similarly to the content-based features since they are most useful when the topic diversity is high.

Features	Judgment		CCAT10		CCAT50		IMDb62	
	FNN	LR	FNN	LR	FNN	LR	FNN	LR
all features	89.43	90.02	75.40	74.20	60.20	60.56	85.25	85.00
(–) Style	**-3.87**	**-4.32**	-3.00	+0.40	-3.40	-2.60	**-6.91**	**-8.39**
(–) Content	-1.43	+0.30	**-3.60**	**-3.00**	**-4.52**	-4.08	-2.77	-2.68
(–) Hybrid	-0.83	-0.29	-3.40	-1.00	-1.28	**-4.68**	-2.02	-5.32

Table 3.6: Feature ablation results

To examine the results further, we generated confusion matrices of the Logistic Regression (LR) classifier applied on the CCAT10 dataset. The effect of removing content-based features is shown in Figure 3.11 where the prediction accuracy of the authors *Alexander Smith* and *Mure Dickie* drops from 96% and 80% (see Figure 3.10) to 84% and 64% respectively. Content-based features are essential in this particular genre (newswire) dataset, since each author usually has different topical interests. For example, among ten authors, *Alexander Smith* mostly discussed topics related to *investment and finance* while topics related to *China* were dominantly written by *Mure Dickie, Benjamin Kang* and *Jane Macartney.* In addition, the writing style between authors in this genre can be very similar. Thus, applying style-based or hybrid features may not be effective. In contrast to CCAT10, removing style-based features in experiments using the Judgment dataset resulted in a greater number of mis-classifications (see Figure 3.12).

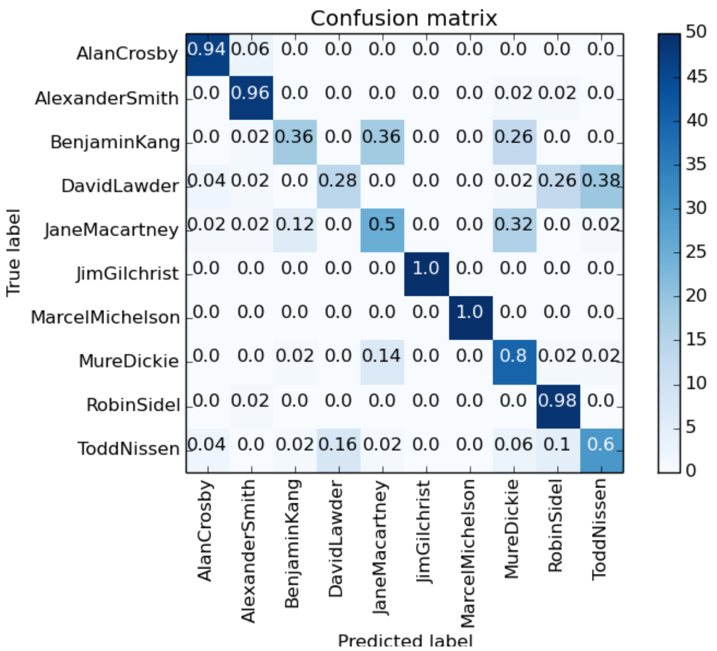


Figure 3.10: Confusion matrix of LR classifier with all features types on CCAT10.

Additional feature exploration was carried out to analyse what types of features are more important to the classifier overall. We performed an analysis using LIME (Ribeiro et al., 2016), a model agnostic framework for intepretability. LIME provides explanations of how a classifier made a prediction by identifying useful input features. We selected a document from each of the datasets and analyzed what kind of features are learned. Figures 3.13, 3.14, 3.15 and 3.16 present the predictions of Logistic Regression (LR) trained on 1000 word unigrams in Judgment, CCAT10, CCAT50, and IMDb62 respectively. In this

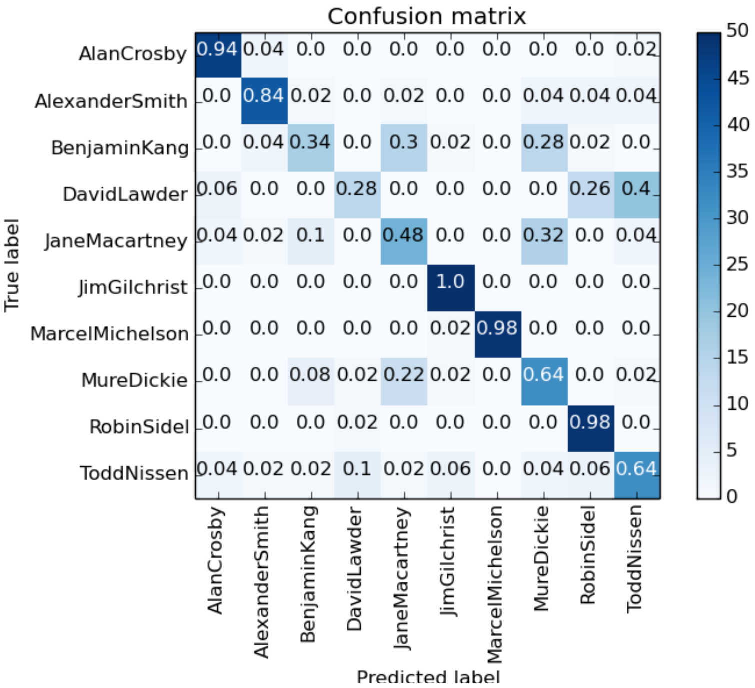
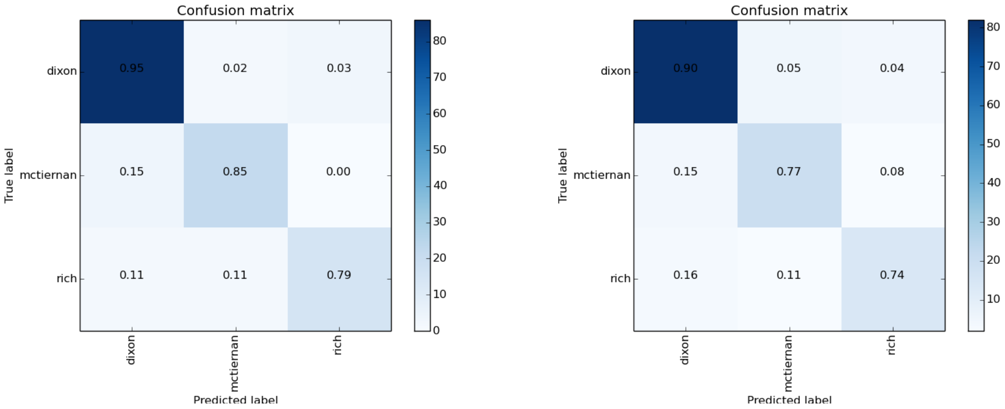


Figure 3.11: Confusion matrix of LR classifier with content-based features excluded on CCAT10.

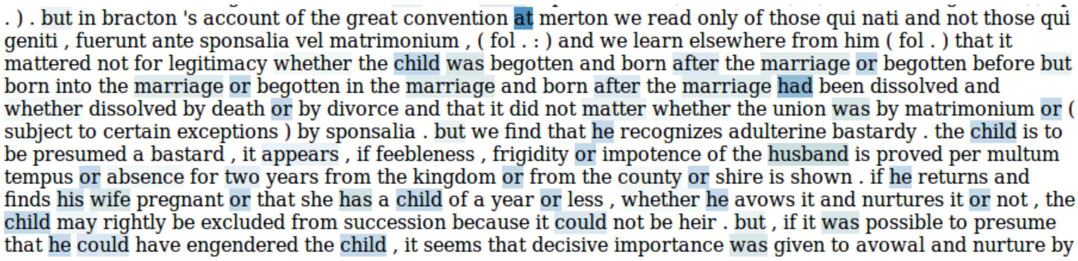


(a) All features

(b) (−) Style

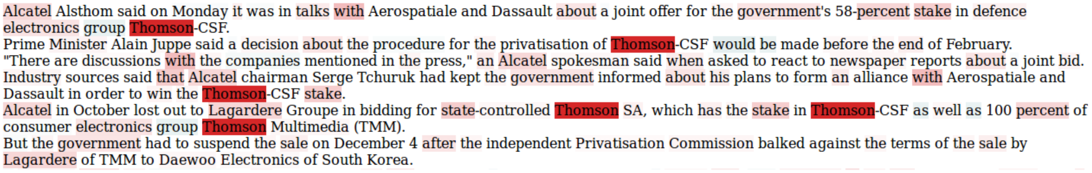
Figure 3.12: Confusion Matrices of LR classifier with different features types on Judgment.

experiment, function words were not removed. For each of the documents presented, LR made correct author predictions with a probability close to 100%. The darker shade indicates how important a particular word is in the attribution decision. We can observe that in CCAT datasets, the classifier put more weight on content-based words such as *Thomson, Canada* and *Toronto*. In contrast to that, function words e.g. *at, had, and, was* appear to be more salient in Judgment and IMDb62.



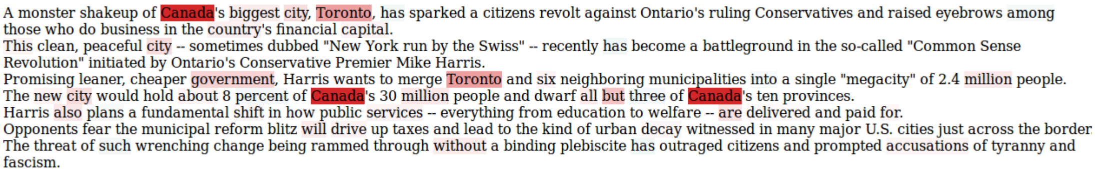
.) . but in bracton 's account of the great convention at merton we read only of those qui nati and not those qui geniti , fuerunt ante sponsalia vel matrimonium , (fol . :) and we learn elsewhere from him (fol .) that it mattered not for legitimacy whether the child was begotten and born after the marriage or begotten before but born into the marriage or begotten in the marriage and born after the marriage had been dissolved and whether dissolved by death or by divorce and that it did not matter whether the union was by matrimonium or (subject to certain exceptions) by sponsalia . but we find that he recognizes adulterine bastardy . the child is to be presumed a bastard , it appears , if feebleness , frigidity or impotence of the husband is proved per multum tempus or absence for two years from the kingdom or from the county or shire is shown . if he returns and finds his wife pregnant or that she has a child of a year or less , whether he avows it and nurtures it or not , the child may rightly be excluded from succession because it could not be heir . but , if it was possible to presume that he could have engendered the child , it seems that decisive importance was given to avowal and nurture by

Figure 3.13: Important word unigrams features in Judgment



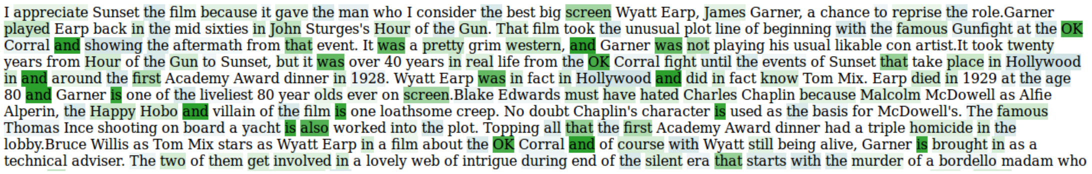
Alcatel Alsthom said on Monday it was in talks with Aerospatiale and Dassault about a joint offer for the government's 58-percent stake in defence electronics group Thomson-CSF.
Prime Minister Alain Juppe said a decision about the procedure for the privatisation of Thomson-CSF would be made before the end of February.
"There are discussions with the companies mentioned in the press," an Alcatel spokesman said when asked to react to newspaper reports about a joint bid.
Industry sources said that Alcatel chairman Serge Tchuruk had kept the government informed about his plans to form an alliance with Aerospatiale and Dassault in order to win the Thomson-CSF stake.
Alcatel in October lost out to Lagardere Groupe in bidding for state-controlled Thomson SA, which has the stake in Thomson-CSF as well as 100 percent of consumer electronics group Thomson Multimedia (TMM).
But the government had to suspend the sale on December 4 after the independent Privatisation Commission balked against the terms of the sale by Lagardere of TMM to Daewoo Electronics of South Korea.

Figure 3.14: Important word unigrams features in CCAT10



A monster shakeup of Canada's biggest city, Toronto, has sparked a citizens revolt against Ontario's ruling Conservatives and raised eyebrows among those who do business in the country's financial capital.
This clean, peaceful city -- sometimes dubbed "New York run by the Swiss" -- recently has become a battleground in the so-called "Common Sense Revolution" initiated by Ontario's Conservative Premier Mike Harris.
Promising leaner, cheaper government, Harris wants to merge Toronto and six neighboring municipalities into a single "megacity" of 2.4 million people.
The new city would hold about 8 percent of Canada's 30 million people and dwarf all but three of Canada's ten provinces.
Harris also plans a fundamental shift in how public services -- everything from education to welfare -- are delivered and paid for.
Opponents fear the municipal reform blitz will drive up taxes and lead to the kind of urban decay witnessed in many major U.S. cities just across the border
The threat of such wrenching change being rammed through without a binding plebiscite has outraged citizens and prompted accusations of tyranny and fascism.

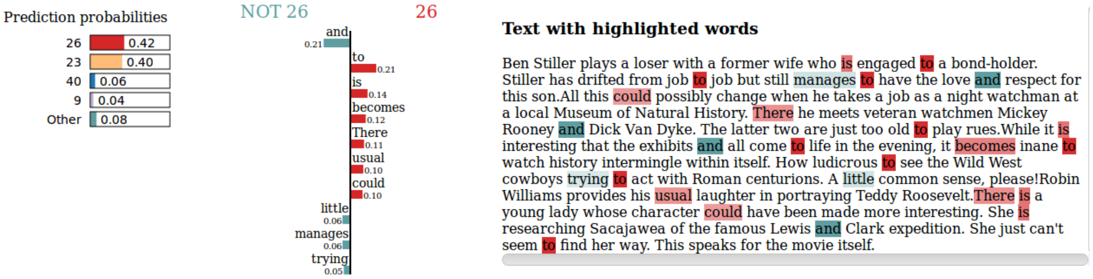
Figure 3.15: Important word unigrams features in CCAT50



I appreciate Sunset the film because it gave the man who I consider the best big screen Wyatt Earp, James Garner, a chance to reprise the role.Garner played Earp back in the mid sixties in John Sturges's Hour of the Gun. That film took the unusual plot line of beginning with the famous Gunfight at the OK Corral and showing the aftermath from that event. It was a pretty grim western, and Garner was not playing his usual likable con artist.It took twenty years from Hour of the Gun to Sunset, but it was over 40 years in real life from the OK Corral fight until the events of Sunset that take place in Hollywood in and around the first Academy Award dinner in 1928. Wyatt Earp was in fact in Hollywood and did in fact know Tom Mix. Earp died in 1929 at the age 80 and Garner is one of the liveliest 80 year olds ever on screen.Blake Edwards must have hated Charles Chaplin because Malcolm McDowell as Alfie Alperin, the Happy Hobo and villain of the film is one loathsome creep. No doubt Chaplin's character is used as the basis for McDowell's. The famous Thomas Ince shooting on board a yacht is also worked into the plot. Topping all that the first Academy Award dinner had a triple homicide in the lobby.Bruce Willis as Tom Mix stars as Wyatt Earp in a film about the OK Corral and of course with Wyatt still being alive, Garner is brought in as a technical adviser. The two of them get involved in a lovely web of intrigue during end of the silent era that starts with the murder of a bordello madam who

Figure 3.16: Important word unigrams features in IMDb62

We also observed a document in the IMDb62 dataset where the classifier assigns similar prediction probabilities to two authors as presented in Figure 3.17. We can notice that the classifier gave the same weight to the function words *and* and *to* which represent two different classes of authors (26 and not 26). The *not 26* represents classes other than author 26. The correct decision of the classifier is more likely helped by the presences of some less significant features such as *is, becomes, There, usual* and *could.*



Text with highlighted words

Ben Stiller plays a loser with a former wife who is engaged to a bond-holder. Stiller has drifted from job to job but still manages to have the love and respect for this son.All this could possibly change when he takes a job as a night watchman at a local Museum of Natural History. There he meets veteran watchmen Mickey Rooney and Dick Van Dyke. The latter two are just too old to play rues.While it is interesting that the exhibits and all come to life in the evening, it becomes inane to watch history intermingle within itself. How ludicrous to see the Wild West cowboys trying to act with Roman centurions. A little common sense, please!Robin Williams provides his usual laughter in portraying Teddy Roosevelt.There is a young lady whose character could have been made more interesting. She is researching Sacajawea of the famous Lewis and Clark expedition. She just can't seem to find her way. This speaks for the movie itself.

Figure 3.17: Explanation of individual predictions of the Logistic Regression classifier on an IMDb62 document using LIME. The bar chart represent the weight given to the most relevant words which are also highlighted in the text.

By applying topic modeling and feature analysis, the most effective features can be predicted for datasets based on their characteristics. Results provide evidence that identifying the topical (or not) nature of the dataset is an important step in determining the best-performing features for a particular authorship attribution problem. Content-based features tend to be suitable for datasets with high topical diversity such as the one constructed from on-line news. While datasets with less topic variances e.g. legal judgment and movie review fit with style-based features.

3.4 Benchmark Work

In addition to the datasets and features analysis, we list previously reported work on the same datasets. In this way, we are able to compare our models against them. In order to ensure a fair comparison, we carefully follow their training and test procedure on conducting our experiments. Table 3.7 describes the previous results of four datasets (Judgment, CCAT10, CCAT50 and IMDb62) using various approaches.

SVM with affix+punctuation 3-grams (Sapkota et al., 2015) This work examined different roles of character n-gram subgroups that correspond to certain linguistic aspects such as morphosyntax, thematic content and style. Sapkota et. al grouped the character n-grams into three categories: affix n-grams, word n-grams and punctuation n-grams. Using SVM as the classifier, their experiment resulted in four datasets including CCAT10 and CCAT50 showing that the combinations of affix and punctuation n-grams are the most effective features among other types of n-grams. They claimed that both of the n-gram types have an ability to capture both morphology and style information which are found to be useful in a single-domain setting.

SVM with 2,500 most frequent 3-grams (Plakias and Stamatatos, 2008) In this approach, a single document was represented as a vector of 2,500 most frequent 3-grams. Then a standard linear SVM model was used as the classifier. In addition to its simplicity, this method was shown to be effective for identifying authors with an accuracy of 80.80% on the CCAT10 dataset.

STM-Asymmetric cross (Plakias and Stamatatos, 2008) Instead of using a vector space model to represent a document, Plakias and Stamatatos proposed a second-order tensor space representation for authorship attribution. In order to handle tensors, they used a generalization of SVM, called Support Tensor Machines (STM) (Cai et al., 2006). The authors claimed that this approach is suitable for cases where only limited training data are available with fewer parameters to be learned. However, their result failed to outperform the standard vector space model by achieving only 78% accuracy on CCAT10.

SVM with bag of local histograms (Escalante et al., 2011) This work proposes local histogram representations over character n-grams for authorship attribution. Using this approach, a set of local histograms are computed across the whole document and are smoothed by kernels centered on different document locations. The representations were claimed to be able to preserve sequential information in the document which may reflect the writing style of the author. This approach obtained the best performance among the other methods on CCAT10 with 86.40% in accuracy. However, our attempt to reproduce their result failed by obtaining only 77% in the accuracy. Another attempt by Potthast et al. (2016a) reported slightly worse accuracy of 75.4%. Thus, we do not consider this work for performance benchmarking in the later chapters.

Token SVM (Seroussi et al., 2013) With only minimal tuning, SVM trained on token frequency features has been known to yield state-of-the-art authorship attribution performance (Koppel et al., 2009). Seroussi et al. used this method as the baseline in their experiments on Judgment and IMDb62 datasets. With relatively large amounts of training data, this method successfully obtained 91.15% and 92.52% in accuracy for Judgment and IMDb62 respectively.

Authorship attribution with topic models (Seroussi et al., 2013) This paper proposes a document representation based on topic modeling. Each document d is represented as a concatenation of two distributions: a document topic distribution and an author topic distribution. In this way, the representation could capture aspects of authorship style while at the same time also represents the authors' interests. Using SVM as the classifier, this method achieved an accuracy of 93.64% for Judgment and 91.79% for IMDb62.

Model	Judgment	CCAT10	CCAT50	IMDb62
SVM with affix+punctuation 3-grams (Sapkota et al., 2015)	-	78.80	**69.30**	-
SVM with 2,500 most frequent 3-grams (Plakias and Stamatatos, 2008)	-	80.80	-	-
STM-Asymmetric cross (Plakias and Stamatatos, 2008)	-	78.00	-	-
SVM with bag of local histogram (Escalante et al., 2011)	-	**86.40**	-	-
Token SVM (Seroussi et al., 2013)	91.15	-	-	**92.52**
Authorship attribution with topic models (Seroussi et al., 2013)	**93.64**	-	-	91.79

Table 3.7: Benchmark work

3.5 Summary

This chapter carried out an analysis of four widely used datasets to explore how different types of feature affect authorship attribution accuracy under varying conditions. The results of the analysis are applied to authorship attribution models based on both discrete and continuous representations. Our experimental results show that our proposed analysis is useful to determine the most effective features based on dataset characteristics. In addition, we also described the detail of previous work that will be used as the benchmark in our experiments.

Chapter 4

Language Models for Authorship Attribution

Language models are a fundamental component in various NLP tasks such as machine translation (Schwenk et al., 2006; Devlin et al., 2014; Luong et al., 2015), image caption generation (Vinyals et al., 2014) and grammatical error correction (Yannakoudakis et al., 2017). This method offers easy access to a large amount of training data and a straightforward learning objective (to predict the next word/character in the sequence) (Rei, 2017). However, only relatively few works have implemented language model for the authorship attribution task. Among them is a work by Peng et al. (2003) who implemented character level Language Modeling (LM) for authorship attribution. The motivation to implement a language model was influenced by some problems faced in the standard attribution model such as language dependency and the difficulty on setting thresholds in feature selection.

In most authorship attribution approaches, feature selection is an important step which may significantly affect task performance. This process usually involves setting a threshold to remove uninformative features (Scott and Matwin,1999). However, defining an optimal threshold can be problematic, because although less useful, rare features can still have an important cumulative effect (Aizawa, 2001). As an example, Stamatatos (2013) demonstrated how the appropriate selection of the number of features is crucial towards the task performance. He examined different sizes of feature sets in intra/cross topic/genre datasets and concluded that each type of dataset has a different optimal feature size. Peng et al. (2003) addressed the problem by avoiding feature selection entirely. In their method, they included all features but use estimation methods from n-gram language modeling to avoid over-fitting a sparse set of training data. They reported that their approach obtained performance higher than 90% accuracy in three datasets which cover three languages: English, Greek and Chinese. However, the datasets involved have a relatively small number of authors (maximum 10 authors). Peng et al. also reported that their result on an academic writing dataset with a more rigid structure only achieved accuracy of 74%, slightly higher than Stamatatos et al. (2000) who implemented feature selection in their approach.

While the results from Peng et al.'s experiments are convincing, n-gram language models consider only a limited context length (Mikolov et al., 2010). This condition prevents the model from acquiring information from longer sequences which might be useful for authorship attribution. More recent work in language modeling achieved state-of-the-art results (Mikolov et al., 2010) by implementing Recurrent Neural Network/Long Short Term Memory (RNN/LSTM) framework which is suitable for processing sequential data. The recurrent connections allow capturing information from arbitrarily long sequences. In addition, the distributed feature representation in neural network architecture allows one to achieve a level of generalization that is not possible with n-gram language models (Mikolov et al., 2013c). A work by Bagnall (2015) is among the few attempts which utilized RNN based language models for authorship attribution. He implemented character level RNN language models for the PAN 2015 Author Verification task (Stamatatos et al., 2015). Bagnall's models successfully obtained the best performance with an average area under the curve (AUC) score greater than 0.8. However, unlike the standard authorship attribution task, the main problem in authorship verification is to decide whether a document of unknown authorship was written by the author of a small set of other documents. Therefore the verification problem focuses on comparing the similarity of two documents rather than capturing an author's writing characteristics.

In this chapter we focus on exploring language modeling for authorship attribution. We begin with presenting the feature selection problem which commonly occurs in standard authorship attribution approaches (Section 4.1). We evaluate two types of features: character and word n-grams and show how the number of features (feature set size) and the value of n influence the task performance. Following this, we apply n-gram language modeling to address the problem (Section 4.2). In contrast to Peng et al. (2003) who only conducted their experiments on a limited type of datasets (in terms of number of authors, genre), we use four different authorship attribution datasets with a range of characteristics in terms of the number of authors, topic/genre and document length: Judgment, CCAT10, CCAT50 and IMDb62 (see details in Section 3.1). Finally, in Section 4.3 we investigate the effectiveness of a neural network model by implementing LSTM-based language modeling for authorship attribution. We discuss how the language model perplexity correlates with the authorship attribution accuracy. Furthermore, we present the limitations of these approaches and possible directions for future work.

4.1 Feature Selection

Character and word n-grams are among the features that have been shown to be effective for authorship attribution (Kjell, 1994; Forsyth and Holmes, 1996; Stamatatos, 2006; Peng et al., 2003; Keselj et al., 2003; Juola, 2004; Stamatatos, 2013; Schwartz et al., 2013). Word n-grams can represent local structure of texts and document topic (Coyotl-Morales et al., 2006; Wang and Manning, 2012) while character n-grams have been shown to be effective for capturing stylistic and morphological information (Koppel et al., 2011; Sapkota et al., 2015). Experiments presented in this section explore the value of n and feature size parameters that might influence the performance of both features in the authorship attribution task.

However, despite the effectiveness it is usually problematic to define an optimal value of n (Peng et al., 2003; Stamatatos, 2009). A small n can be inadequate to capture sufficient information, while a large n will significantly increase the dimensionality of the representation and create sparse training data. In addition, the optimal n-value is usually language dependent, since average word length varies across languages.

4.1.1 Experimental setup

The experiments in this section use the Support Vector Machine (SVM) implementation from Scikit Learn (Pedregosa et al., 2011). We followed previous work by using the provided train/test partitions for both CCAT datasets and the 10-fold cross validation for Judgment and IMDb62. The SVM hyper-parameters were fixed for all the experiments. We used a Radial-basis function (RBF) kernel with the values of C and gamma set to 10.0 and 0.0001. We conducted experiments with two types of features: character and word n-grams. A document is represented as a frequency vector of the respective features. We performed an experiment with various values of n and feature sizes. We did not apply any text pre-processing to the documents.

4.1.2 Value of n

This experiment was carried out to see the effect of different values of n on both character and word n-grams on authorship attribution performance. In this experiment, we set the feature set size to the 100 most common n-grams when varying the value of n from 2 to 10 for character n-grams and 1 to 5 for word n-grams. Figure 4.1 and 4.3 demonstrate the accuracy obtained using various values of character and word n-grams in all four datasets. The x-axis represents the value of n while the y-axis indicates the accuracy obtained. From the figure, it is noticeable that the performances of both word and character n-grams are affected by the choice of n. For models with character n-grams, it is obvious that there is a sharp increase in accuracy when 3-grams are used, especially for the CCAT10, CCAT50 and IMDb62 datasets as shown in Figure 4.1. This result is in agreement with previous work (Stamatatos, 2013; Sapkota et al., 2015) which used character 3-gram features in their experiments.

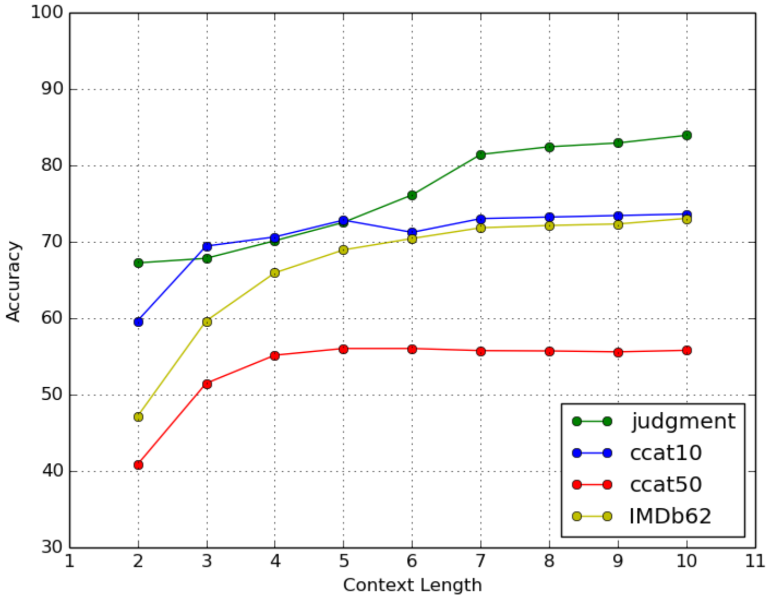


Figure 4.1: The accuracy obtained with different values of character n-grams

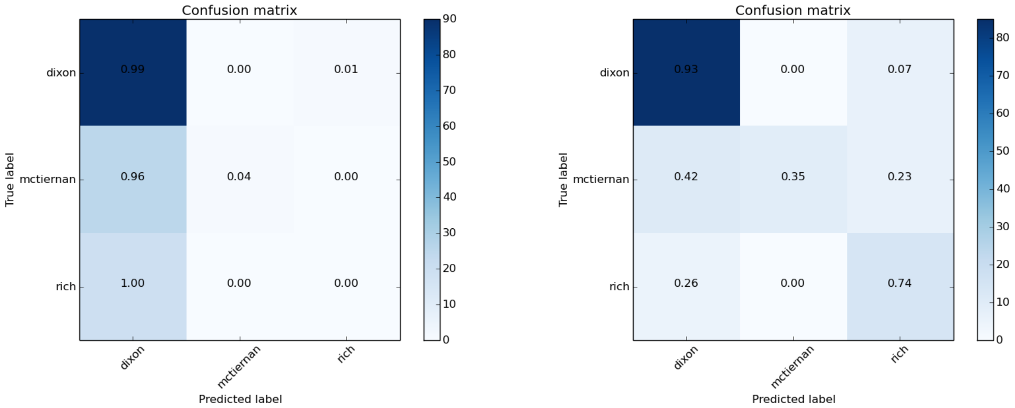
Table 4.1 shows the list of 10-most common character 3-grams for each dataset. In Judgment, CCAT10 and CCAT50, the character 3-grams are dominated by n-grams which are part of the word. It can be either the first/last three or the mid characters of the word. For example, in CCAT10 and CCAT50 the 3-grams `ill` is mostly from the word *billion* or *million* while 3-grams `_ch` is derived from *China* or *Chinese*. In contrast to that, in IMDb62 most of the common 3-grams are stopwords e.g. *a, an, of, to, and, is.* This might be affected by the informal genre of the dataset which influenced the author's choice of words. Compared to three other datasets which discuss more formal topics, IMDb62 has the shortest average word length. Results from this experiment explain how the character 3-grams could capture both topic and style, while in the same time support the findings from the previous work (Sapkota et al., 2015). Sapkota et al. performed a thorough analysis of the effectiveness of character 3-grams. They divided the character 3-grams into three groups: affix, word and punctuation n-grams and observed that each group captures different information which covers both topic and style.

Judgment	CCAT10	CCAT50	IMDb62
`_¨_`	`_ca`	`_bu`	`_a_`
`_(_`	`_fr`	`_ca`	`_an`
`_)_`	`_ch`	`_ch`	`_of`
`_at`	`con`	`_ne`	`_to`
`_ex`	`ear`	`_wa`	`and`
`_pa`	`ill`	`are`	`he_`
`ain`	`men`	`ce_`	`ing`
`com`	`pro`	`ear`	`is_`
`int`	`rs_`	`hat`	`nd_`
`was`	`ted`	`per`	`ng_`

Table 4.1: 10 most common character 3-grams for each dataset

From Figure 4.1 we can observe that performance can still be improved by adding more n-grams. For CCAT10, CCAT50 and IMDb62, the authorship attribution model obtained the best performance by using up to character 5-grams before the graph starts to plateau or even decline. An exception is found in the Judgment dataset in which character 3-grams do not really help increase the accuracy compared to 2-grams. Unlike the other datasets, the Judgment graph shows an upward trend before reaching a peak at 7-grams and starts to level off. The confusion matrices in Figure 4.2 clearly show how the classifier could identify the authors better when using up to character 7-grams. By adding more n-grams, the model minimized the error on identifying the author *Rich* by obtaining 74% in accuracy.

A similar trend is found for models with word n-grams as shown in Figure 4.3. The performances of all the datasets except Judgment increased up to



(a) Char up to 3-grams

(b) Char up to 7-grams

Figure 4.2: Confusion matrices of Judgment with different values of n

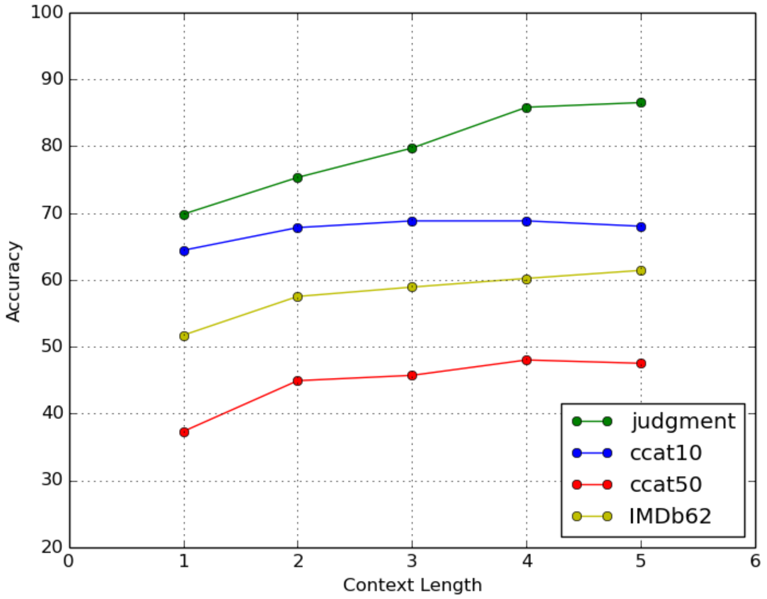
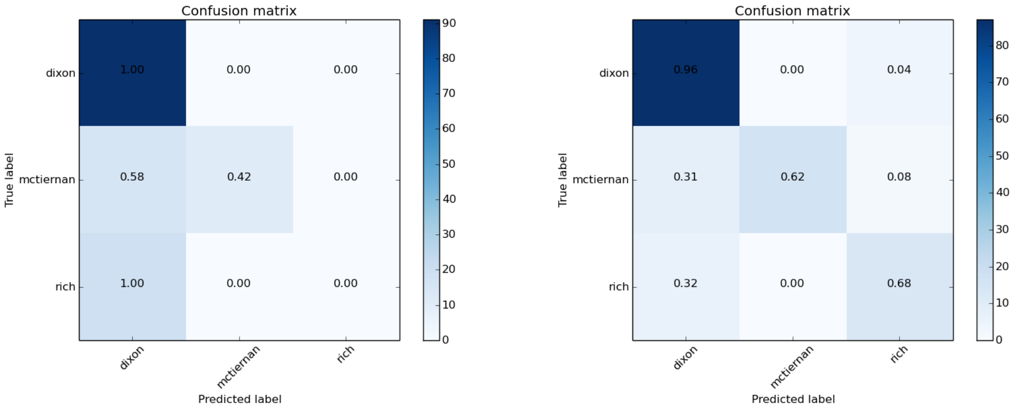


Figure 4.3: The accuracy obtained with different values of word *n*-grams

word bi-grams and remained flat after that. Since we were not performing any text pre-processing, the 10-most common word bi-grams are dominated by the combination of stopwords such as *by the, for the, and the*, etc (see Table 4.2). Some content words like *hong kong, the company, the movie, the film* were also captured. On the other hand, Judgment reached its optimal performance by using up to word 4-grams. We compared two confusion matrices of Judgment which were generated with different set of features and discovered that models with word 4-grams are better at distinguishing authors (see Figure 4.4). In Judgment, word bi-grams failed to represent the unique writing style of each author. The cause can be explained with this following example: the word bi-gram *in the* is found 3401 times in McTiernan, 1332 times in Rich and 20082 times in Dixon writings, while the four-gram *in the present case* was only used 61 times by McTiernan, 26 times by Rich and 532 times by Dixon. The frequency ratio of both n-grams for each authors might be similar (since bi-gram *in the* is a subset of 4-gram *in the present case*). However, we can see that the use of *in the present case* is not as common as the use of *in the*, which make it a strong indicator of unique writing style for a particular author.

Judgment	CCAT10	CCAT50	IMDb62
and_the	by_the	for_the	and_the
at_the	for_the	he_said	in_the
by_the	he_said	hong_kong	of_the
for_the	in_the	in_the	on_the
from_the	of_the	of_the	one_of
it_is	on_the	on_the	the_film
on_the	said_the	said_the	the_movie
to_be	to_be	the_company	to_be
upon_the	to_the	to_be	to_the
which_the	with_the	to_the	with_the

Table 4.2: 10 most common word bi-grams for each dataset



(a) Word up to bi-grams

(b) Word up to 4-grams

Figure 4.4: Confusion matrices of Judgment with different value of n

Through the experiments conducted, we have shown the problems of choosing the optimal value of n. Although previous work has suggested certain optimal values, we have shown that they may not be suitable for all datasets. We argue that there are two main factors that are responsible for the performance difference in Judgment. First, among other datasets, Judgment has an imbalanced number of documents per author. Dixon produced almost three and five times as many documents as McTiernan and Rich. This caused misclassification of the documents produced by the minority class as have been shown in Figure 4.2a and 4.4a. Second, compared to the three other datasets, Judgment has the highest topical similarity (see discussion in Chapter 3). Furthermore, due to the genre of the dataset (legal judgment), the writing style between authors tends to be similar. Thus, we believe that the optimal values suggested in previous work (character 3-grams and word bi-grams) are not suitable for this dataset.

4.1.3 Influence of Feature Set Size

Another factor that influences the accuracy of authorship attribution is the number of features used. We performed experiments with a range of feature sizes from 500 to 4000 and 100 to 1000 for the character and word-based models respectively. 3-grams were used for character-based model and up to bi-grams for the word-based model. Figures 4.5 and 4.6 show the accuracy obtained for each dataset with various numbers of features. For character n-grams using the 2,500 most frequent 3-grams produced optimal performance before the accuracy started to plateau (Figure 4.5). Similar trends are shown for word n-grams (Figure 4.6). Optimal accuracy was reached by using 600 features.

The main observation is that, unlike our previous experiments with the value of n, all four datasets tended to have similar thresholds for the feature size. As mentioned previously, Stamatatos (2013) conducted experiments with intra/cross topic/genre attributions and found that the thresholds for optimal performance differ for each type of tasks. In the intra topic/genre task, the training and test sets belong to the same genre and thematic area, while in the cross topic/genre, the topic/genre may be different for training and test sets. In our experiment, the task is considered as intra topic/genre since the datasets involved have the same genre and topic for both training and test data. We found that, our results on the character based model are consistent with Stamatatos' results in the intra topic/genre task. Both of our results show that 2500 features are optimal for the model with character 3-grams.

Nevertheless, we have shown that feature selection is crucial. The condition of the datasets such as unbalanced data, type of the task (cross/intra attribution) are some of the factors that need to be considered.

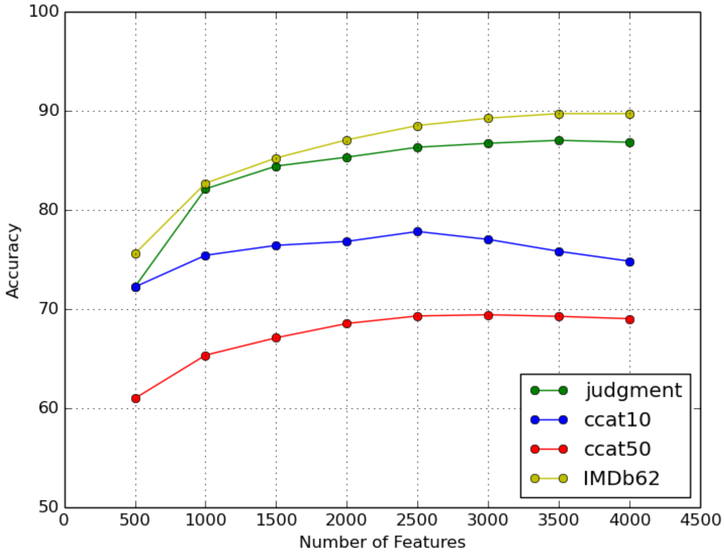


Figure 4.5: Authorship attribution accuracy with different numbers of character tri-grams

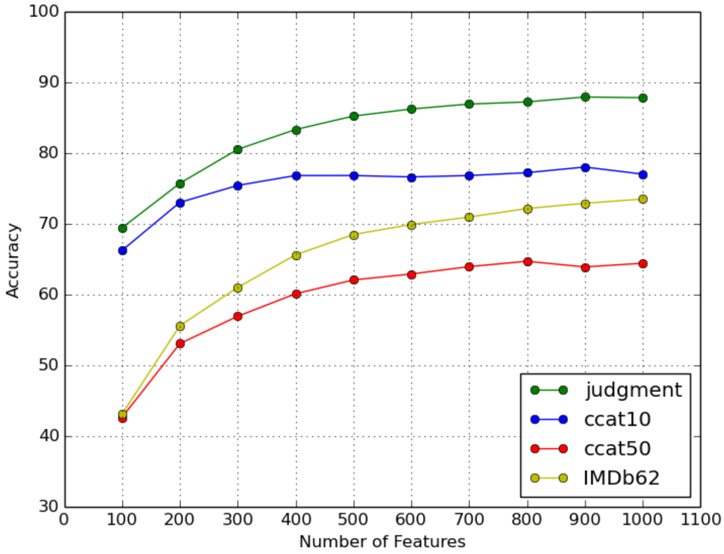


Figure 4.6: Authorship attribution accuracy with different numbers of word bi-grams

4.2 *N*-gram based Language Modeling for Authorship Attribution

In this section, we present an n-gram based language modeling approach which attempts to address the problem of selecting suitable features, as outlined in the previous section. We replicate Peng et al.'s (2003) experiments with more diverse characteristics of datasets in terms of the number of authors, topic/genre and document length. Compared to Peng et al., who only used datasets with a maximum of 10 authors, the datasets involved in our experiments have a wider range of numbers of authors (see details in Table 3.1). Furthermore, our datasets also have different levels of topical diversity. Thus, we can examine whether the n-gram based language modeling is effective for a wider range of datasets.

4.2.1 Overview of *N*-gram Language Modeling

Jurafsky and Martin (2000) describe Language Models (LM) as models that assign probabilities to sequences of words. This probabilities are essential in many NLP tasks e.g speech recognition, spelling correction and machine translation (Schwenk et al., 2006; Devlin et al., 2014; Luong et al., 2015; Vinyals et al., 2014; Yannakoudakis et al., 2017). These tasks were similar in that they used on using probabilities of word sequences to find the most probable solution. As an example in spelling correction, assume that we need to find and correct spelling errors in this following sentence *Their like to play football together* in which *They* was incorrectly typed as *Their.* Supposing training is effective, using language models we can easily spot the error as the phrase *They like* is more probable than *Their like.* The error is not limited to spelling errors but also grammatical errors.

Given the word sequence $W = w_1, w_2, ..., w_N$, and let w_1^{n-1} be a sequence of preceding words of w_n, n-gram language models work by predicting the probability of the sequences $P(w_1, w_2, ..., w_N)$ using the chain rule of probability:

$$\begin{aligned} P(w_1...w_N) &= P(w_1)P(w_2|w_1)P(w_3|w_1^2)...P(w_n|w_1^{n-1}) \\ &= \prod_{k-1}^{N} P(w_k|w_1^{k-1}) \end{aligned} \tag{4.1}$$

However, using the n-grams model, the probability of the next word can be approximated by just the last few words. For example in the bi-gram model, instead of calculating the probability

$$P(\texttt{story|A mixture of truth and fiction! Okay if you know the true}) \tag{4.2}$$

can be approximated with the probability

$$P(\texttt{story|true}) \tag{4.3}$$

Using the Markov assumption, we do not need to use the entire history. The n-gram probabilities approximation of a complete word sequence can be computed by substituting Equation 4.1 with:

$$P(w_1^n) \approx \prod_{k=1}^{n} P(w_k|w_{k-1}) \tag{4.4}$$

Then, to estimate a particular bi-gram probability of a word w_n given a previous word w_{n-1}, we can use Maximum Likelihood Estimation (MLE). The MLE estimate can be computed by counts of the bigram $C(w_{n-1}w_n)$ and normalized by the sum of all the bi-grams which contain the same first word w_{n-1}:

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)}{\sum_w C(w_{n-1}w)} \tag{4.5}$$

Equation. 4.5 can be simplified by substituting the denominator with the unigram counts of the word w_{n-1}

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)}{C(w_{n-1})} \tag{4.6}$$

A major drawback of MLE is the poor estimate of zero or low frequency counts of n-grams which are more common in small training sets. To address the problem, *smoothing* can be applied to modify the probability of n-grams. Some non-zero counts are discounted/lowered in order to get the probability mass that will be assigned to the zero counts. Among several smoothing techniques, Kneser-Ney (Ney et al., 1994) is one of the most common methods. It works by re-estimated count c^* by subtracting a fixed discount D from each count while in the same time handling the backoff distribution. Assuming a proper coefficient α on the backoff, the probability of n-grams with Kneser Ney smoothing can be formalised as follows:

$$P_{KN}(w_i|w_{i-1}) = \begin{cases} \frac{C(w_{i-1})-\mathbf{D}}{C(w_{i-1})}, & \text{if } C(w_{i-1}w_i) > 0 \\ \\ \alpha(w_i)\frac{|\{w_{i-1}:C(w_{i-1}w_i)>0\}|}{\sum_{w_i}|\{w_{i-1}:C(w_{i-1}w_i)>0\}|} & \text{otherwise.} \end{cases} \tag{4.7}$$

For the experiment in this section, we used a statistical n-gram language model provided by the publicly available toolkit SRILM (Stolcke, 2002)[1]. SRILM implements both original and modified Kneser-Ney discountings (Chen and Goodman, 1996). In modified Kneser-Ney discountings, for each n-gram order, it uses three discounting constants, one for one-count n-grams (n_1), one for two-count n-grams (n_2), and one for three-plus-count n-grams (n_3). The discounting constants can be computed as follows:

$$\begin{aligned} Y &= \frac{1}{(n_1 + 2 * n_2)} \\ D1 &= 1 - 2Y(\frac{n_2}{n_1}) \\ D2 &= 2 - 3Y(\frac{n_3}{n_2}) \\ D3+ &= 3 - 4Y(\frac{n_4}{n_3}) \end{aligned} \tag{4.8}$$

Finally, to evaluate the model, we can use *perplexity*:

$$\begin{aligned} PP &= P(w_1 w_2 ... w_N)^{-\frac{1}{N}} \\ &= \sqrt[N]{\frac{1}{P(w_1 w_2 ... w_N)}} \end{aligned} \tag{4.9}$$

4.2.2 *N*-gram Language modeling for Authorship Attribution

We followed Peng et al. (2003) by building separate language models for each of the authors. In this experiment, we built both word and character level models. Figure 4.7 shows the flow of creating a language model for each author using SRILM. The lexicon was built by listing all n-gram words or characters found in the training set from all of the authors. This will ensure that the same dictionary is used when building language models across authors.

An author class $a \in A = \{a_1, ...a_n\}$ is assigned to a new document d if the language model of author a assigns the lowest perplexity given the document d.

$$a^* = argmin_{a \in A}\{PPL(d|a)\} \tag{4.10}$$

4.2.3 Results and Discussion

We performed experiments with a range of n-gram lengths for both word and character-level models. Table 4.3 describes the average number of characters and

[1] https://www.sri.com

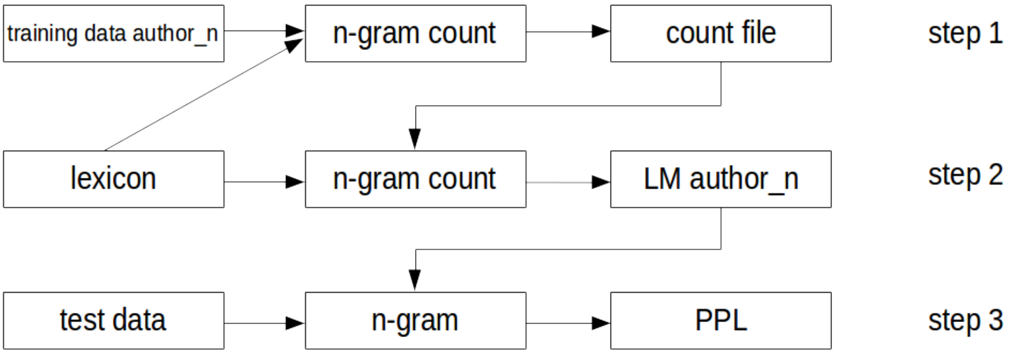


Figure 4.7: Creating LM using SRILM (Chen, 2014)

Dataset	Training		Test	
	#char	#word	#char	#word
Judgment (Dixon)	9.7M	1.9M	1.1M	213K
Judgment (McTiernan)	2.7M	539K	303K	60K
Judgment (Rich)	2M	398K	224K	44K
CCAT10	154.5K	29K	154.5K	29K
CCAT50	153K	29.2K	153K	29.2K
IMDb62	1.5M	311K	168K	34.5K

Table 4.3: Average number of characters and words per author

words per author for the training and test sets. In this experiment, we used 10% of the training data as the development set. For Judgment, we provide details per author (Dixon, McTiernan, Rich) since this dataset has an imbalanced number of documents per author. Figures 4.8 and 4.9 show the authorship attribution accuracies obtained with different n-gram lengths on the development set using character and word-level models respectively. The best accuracy for the character-level model was achieved by using either character tri-grams or four-grams (see Figure 4.8). While in the word-level model, word bi-grams are found to be the most effective.

Compared to the word-level model, the character model produced better performances in all datasets. For example in IMDb62, the model obtained the best accuracy of 87.45% in contrast to the word-level model which only achieved 65.75%. However, in the Judgment dataset, the word-level model obtained 80.50%, almost similar to the character model with 81.82%. We examined the

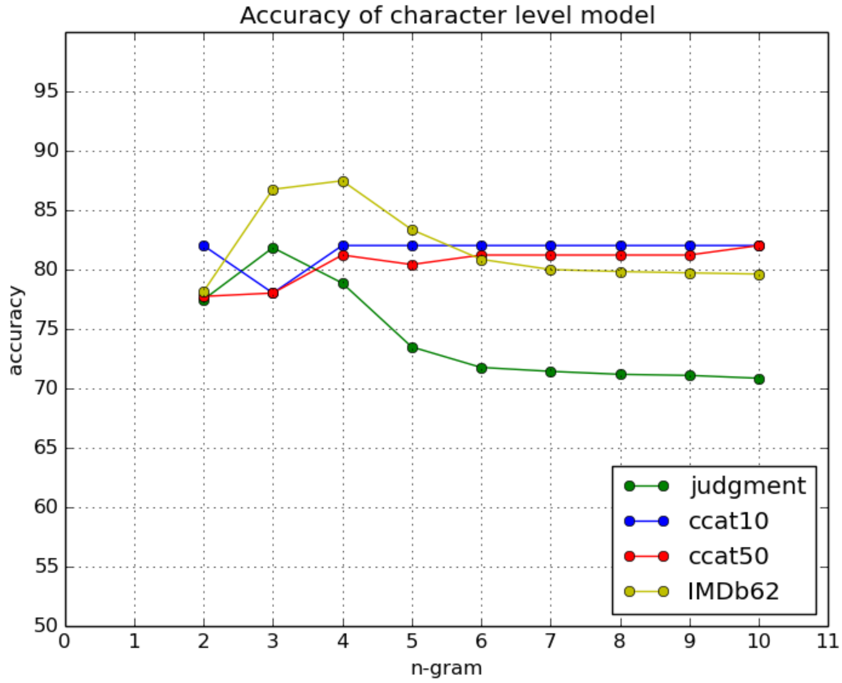


Figure 4.8: Authorship attribution performance using character-level language model in development set

cause of performance differences by evaluating the quality of language models independently from the authorship attribution task. Figure 4.10 and 4.11 show the average perplexity of character and word-level models for different n-gram lengths. The results clearly demonstrate that the word-level language models have significantly higher perplexity compared to the character-level model. The poor performances of word-level models are likely due to the small training set size. This argument is supported by the fact that the word-model still produces good performance on the larger dataset (Judgment). It is interesting that by representing documents at the character level, we can have more training data in which could provide better quality of language models. In addition, the vocabulary size of character-level models which is smaller than the vocabulary size of word models helps to reduce the sparse data problem which might be encountered in the experiments.

Table 4.4 compares the results of the approach reported here against previous authorship attribution work that used the same datasets. The results presented are the accuracy obtained in the test set. Most of the previous approaches presented in the table, used SVM with various feature types (see Section 3.4). Language modeling-based approaches failed to provide more accurate predictions by

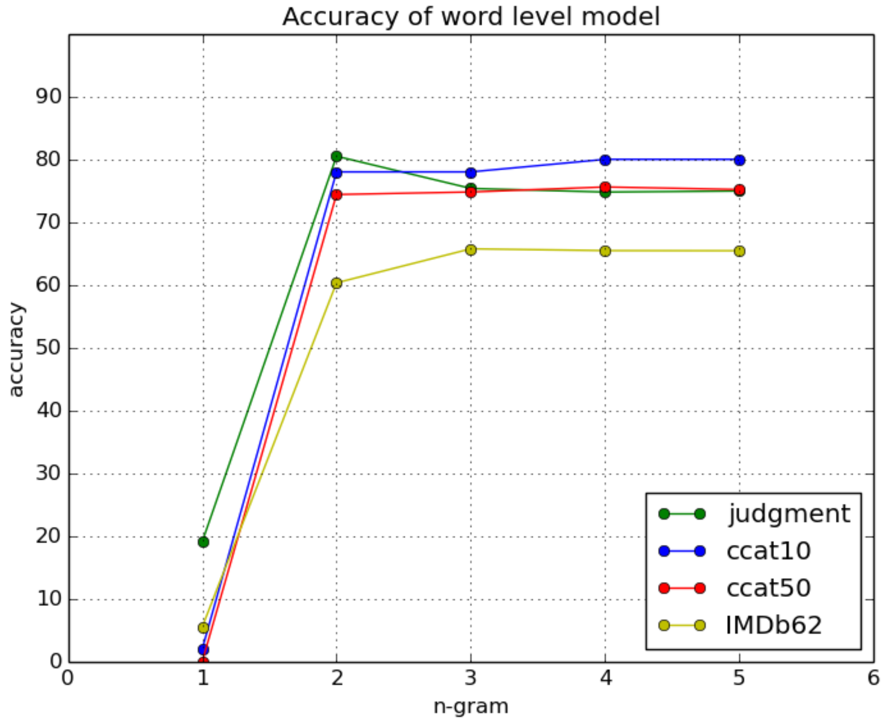


Figure 4.9: Authorship attribution performance using word-level language model in development set

obtaining lower accuracy than the previous results in all four datasets. We argue that there are two main underlying factors causing the poor performance. First, the poor quality of language model due to the small training size. Evidence to support this argument can be seen from the results of the character-level model which gained better performance than word-level model.

Second, we found that the language model-based approach is more suitable for datasets with higher topical diversity and/or distinct idiosyncrasies in writing style. Among the four datasets, CCAT10 and CCAT50 obtained better accuracies relative to the previous results. In our previous experiments (Chapter 3), we have shown that both CCAT datasets have higher topical diversity compared to Judgment and IMDb62. This argument is also supported by Peng et al. (2003) who obtained very high performance (more than 90% accuracy) for all datasets except for the dataset with a more rigid structure or uniform writing style (i.e. academic writing). Peng et al. reported that the datasets used in their experiments (especially for English and Chinese languages) have distinct writing styles since they were constructed from novels written by several famous authors such as Charles Dickens and Shakespeare.

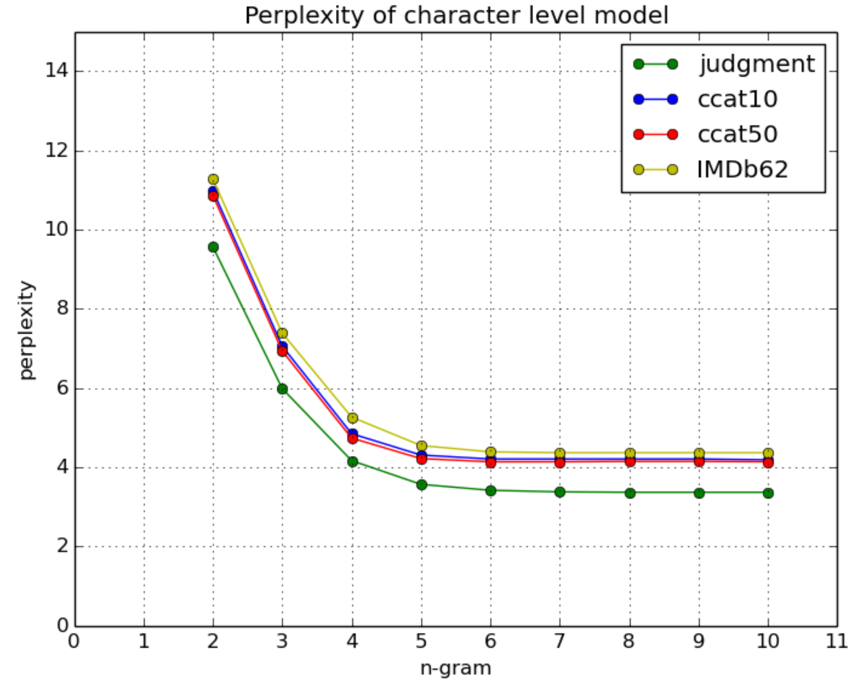


Figure 4.10: Perplexity of character-level language model in the development sets.

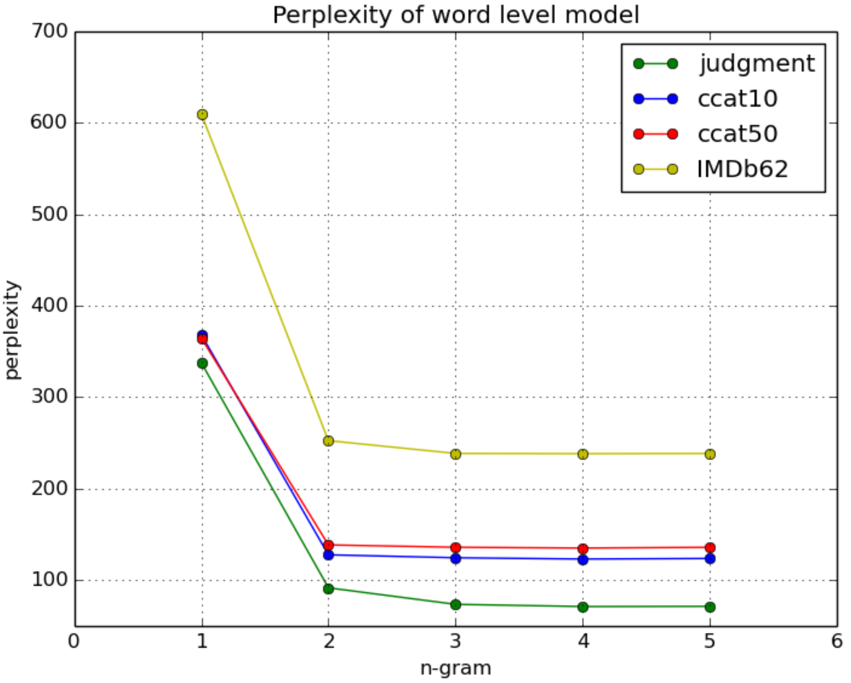


Figure 4.11: Perplexity of word-level language model in the development sets.

Model	Judgment	CCAT10	CCAT50	IMDb62
SVM with affix+punctuation 3-grams (Sapkota et al., 2015)	-	78.80	**69.30**	-
SVM with 2,500 most frequent 3-grams (Plakias and Stamatatos, 2008)	-	**80.80**	-	-
Token SVM (Seroussi et al., 2013)	91.15	-	-	**92.52**
Authorship attribution with topic models (Seroussi et al., 2013)	**93.64**	-	-	91.79
Character-level n-gram LM	80.78	75.00	64.76	87.39
Word-level n-gram LM	80.55	59.60	46.52	68.00

Table 4.4: Comparison against previous results.

Nevertheless, apart from the low attribution performances, we found that the language model-based approaches can be used to address problems faced in the feature selection. However, due to the sparse data problem, it limits the model for getting longer-distance information which would probably be useful for authorship attribution. As previously demonstrated in Figures 4.10 and 4.11, a longer context did not help to improve the perplexity. In the next section, we aim to address the drawbacks of n-gram language models by using the Long Short Term Memory (LSTM) model which has the ability to preserve information from long context sequences.

4.3 LSTM-based Language Model for Authorship Attribution

In this section, we apply the Long Short Term Memory (LSTM)-based language model to the authorship attribution task. We examine whether information from longer contexts is useful for authorship attribution. Unlike the n-gram language models, where only a limited context length would be considered, LSTM allows conditioning the model on all previous words/characters in the document (Mikolov et al., 2010).

We begin this section with a short overview of Recurrent Neural Networks

(RNNs). RNN (Rumelhart et al., 1986) is a type of neural network for processing sequential data. This architecture can be considered as an unfolded feed-forward neural network with a single shared model which operates on all time steps and all sequence lengths (Goodfellow et al., 2016). Figure 4.12 illustrates the computational graph of an RNN. At each time step t, the model has x_t, h^t, o^t, y^t and L^t which represent input, hidden layer activation, output, target and loss. Training the model starts with forward propagation which applies these computations at each time step:

$$\begin{aligned} h^{(t)} &= tanh(Wh^{(t-1)} + Ux^{(t)} + b) \\ o^{(t)} &= Vh^{(t)} + c \end{aligned} \tag{4.11}$$

where U, W and V are the weight matrices between input to hidden, hidden to hidden and hidden to output respectively and b, c are bias vectors.

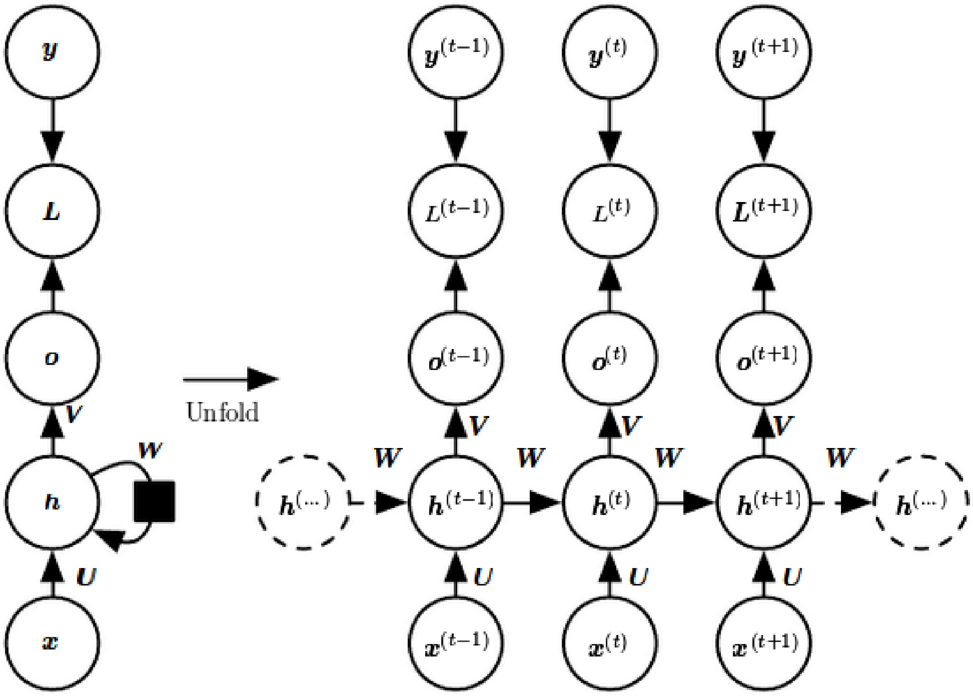


Figure 4.12: Recurrent Neural Network (RNN) architecture (Goodfellow et al., 2016)

To obtain target $y^{(t)}$, a softmax operation can be applied over the output $o^{(t)}$

$$\hat{y}^{(t)} = softmax(Vh_t) \tag{4.12}$$

Then we can use the negative log-likelihood to compute the loss/error between the predicted value and the correct value in each step. For the total loss given a sequence of x values paired with a sequence of y values:

$$\begin{aligned} L(\{x^{(1)},...,x^{(T)}\},\{y^{(1)},...,y^{(T)}\}) &= -\sum_t L^{(t)} \\ &= -\sum_t logp_{model}(y^{(t)}|\{x^{(1)},...,x^{(t)}\}) \end{aligned} \tag{4.13}$$

The gradient of this loss function can be computed by performing backpropagation through time (BPPT) (Werbos, 1990) which involves a forward propagation from left to right followed by a backward propagation from right to left of the unrolled graph. RNN models have been applied for many NLP tasks including language modeling, speech recognition and machine translation (Mikolov et al., 2010; Mikolov and Zweig, 2012) and are reported to bring improvement to the performance.

Theoretically, RNNs can learn information from arbitrarily long sequences. However, in practice it suffers from the vanishing gradient problem (Hochreiter, 1998; Bengio et al., 1994). Consider a language model trying to predict the last words of these following sentences:

`The chef cooks in the` ***`kitchen`***

`Since April 1995, when the yen hit a high of 80 to the dollar, the Japanese currency has weakened. The rate is now about 114, meaning it takes more yen to buy one` *`dollar`*`.`

In theory, RNN should be able to predict correctly the last words in both sentences. However, the contribution of gradient values during the back-propagation phase gradually vanishes. This problem is more likely to occur in long sentences where the gap between relevant information and the point where it is needed is large. Thus in the latter sentence, the probability that *dollar* would be predicted correctly is smaller than the word *kitchen* in the first sentence. To address the problem, Hochreiter and Schmidhuber (1997) proposed Long Short Term Memory (LSTM), a special type of RNN which is capable of learning long-term dependencies. LSTM addresses the problem by introducing a memory cell with gating units in its architecture. The gating units have the ability to control whether information from previous states need to be removed or preserved, meaning that the vanishing gradient problem can be avoided.

4.3.1 Character-level Language Models with LSTM

We limit the implementation of the LSTM-based model to the character-level since the datasets available are not large enough to train word-based models. In language modeling, given a sequence of N characters $C_1, ..., C_N$, the sequence probability can be calculated using Equation 4.4 by substituting the words with characters. In the LSTM-based model, the probability can be estimated by feeding the encoded character input vector x_t, the previous hidden state h_{t-1}, and the previous memory cell c_{t-1} into the LSTM one at a time. The next hidden state h_t then can be produced via the following calculation:

$$\begin{aligned}
i_t &= \sigma(W^{(i)}x_t + U^{(i)}h_{t-1} + b^{(i)}) \\
f_t &= \sigma(W^{(f)}x_t + U^{(f)}h_{t-1} + b^{(f)}) \\
o_t &= \sigma(W^{(o)}x_t + U^{(o)}h_{t-1} + b^{(o)}) \\
u_t &= tanh(W^{(u)}x_t + U^{(u)}h_{t-1} + b^{(u)}) \\
c_t &= i_t \odot u_t + f_t \odot c_{t-1} \\
h_t &= o_t \odot tanh(c_t)
\end{aligned} \tag{4.14}$$

The LSTM cell has three gates: an *input gate* i_t, a *forget gate* f_t and an *output gate* o_t; a *memory cell* c_t and a hidden state h_t. $W^{(*)}$, $U^{(*)}$, $b^{(*)}$ denote the weight matrix between hidden to hidden, input to hidden and bias in each gate. A number between 0 and 1 is produced in each gate, which represents how much information should be kept or removed. All of the gates receive input x_t from the current time step t and previous hidden state h_{t-1}. The forget gate will decide to what extent information from the previous state is forgotten, while the input gate controls what new information will be added to the cell state. The output gate filters the exposure of the internal memory cell of the current state. A partial view of the internal memory cell c_t is represented by the hidden state h_t. Finally, the target word at time step t and the total loss can be predicted by Equations 4.12 and 4.13 respectively.

Figure 4.13 presents a working example of a character-level LSTM-based language model. Suppose we want to train the LSTM on the training sequence "cat_" (_ denotes space). First, each character will be encoded using 1-of-k encoding, where k is the size of vocabulary. Given the target character for each time step, the LSTM will be trained to assign the maximum probability. Training is performed using the back-propagation through time algorithm and is repeated until the network converges and its predictions are consistent with the given labels.

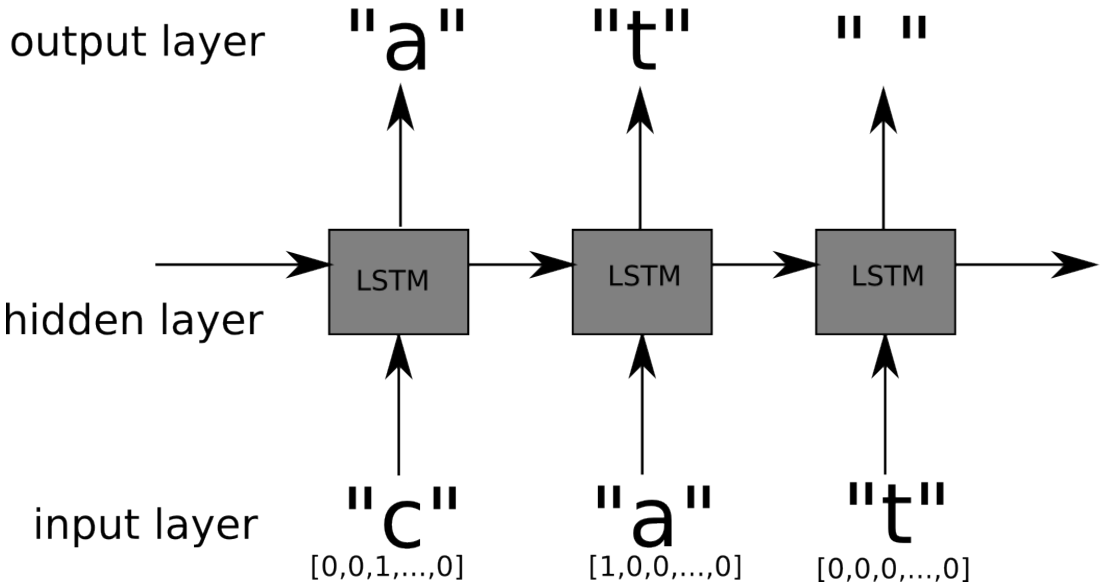


Figure 4.13: An example of character-level LSTM language models

If $C_1^T = [C_1, ..., C_T]$ is the sequence of characters in the training corpus, then training involves minimizing the negative log-likelihood (NLL) of the sequence

$$NLL = -\sum_{t=1}^{T} logPr(C_t|C_1^{t-1}) \tag{4.15}$$

and the perplexity (PPL) of a language model over the character sequence is calculated by

$$PPL = exp(\frac{NLL}{T}) \tag{4.16}$$

Finally, given a document d and a fixed set of candidate authors $a \in A = \{a_1, a_2, ..., a_m\}$, a separate character-level language model for each of the authors. To categorize a new document d, we pick the language model of an author a that has the lowest perplexity, see Equation 4.10.

4.3.2 Experiment

We performed experiments with the datasets mentioned in Table 4.3. For each of the datasets, we used 10% of the training data as the validation set. Our LSTM-based language model consists of a single LSTM hidden layer with 128 hidden units. In total there are 110,918 parameters in our model. This model is considered small compared to previous work (Mikolov et al., 2010; Zaremba et al., 2014). We did not implement a larger model since our training data per author is limited and fairly small for the language modeling task. Each character in the document is represented using one hot encoding (Bishop, 2006). We set a fixed vocabulary list consisting of 70 characters including the characters of the 26 English alphabet, 10 digits and 34 other characters:

a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v,
w, x, y, z, 1, 2, 3, 4, 5, 6, 7, 8, 9, 0, -, ,’, ;, ., !, ?, :, ",
‘, /,\,|,_,@,#,$,%, ˜,&, *,;̃ ’, +, =, <, >, (,), [,], , , \n, "
"

Optimization
The model was trained using truncated backpropagation through time (Williams and Peng, 1990) and optimized using the Adam update rule (Kingma and Ba, 2014). The learning rate was set to 0.002 with decay rate 0.95. We conducted hyper-parameter tuning including batch size and dropout rate (from hidden to output layer) on the CCAT10 validation set. We found that 5 and 0.75 are the optimal values for batch size and dropout rate respectively. Gradients were averaged over each batch. For CCAT10, CCAT50 and IMDb62, we trained each language model of the authors for 100 epochs, while for Judgment, we set 20 epochs for Dixon and 100 epochs for the rest of the authors. Note that Dixon has an almost 10 times larger training size compared to the others authors. We found that the chosen number of epochs were sufficient for the model to reach convergence. Finally, we picked only the best model on the validation and used it to perform evaluation on the test set.

4.3.3 Results and Analysis

In this section we present results on four datasets using the model described in Section 4.3.2. First, we examined the effect of the number of time steps in truncated backpropagation through time (TBPTT) to the language model perplexity and how it correlates with the authorship attribution performance. Then we provide an analysis of the limitations LSTM-based language model for authorship attribution

4.3.4 The effect of perplexity on authorship attribution performance

It is common practice to perform truncated backpropagation (Williams and Peng, 1990) for training LSTM/RNN. Sutskever (2013) explained that the main problem of BPTT is the high cost of parameter update which limits the use of large numbers of iterations. He reported that the cost of gradient update of an RNN/LSTM on a sequence of length N is equal to the cost of a forward and a backward pass in a neural network model with N layers. Considering the document length of the datasets involved in our experiment, applying truncated backpropagation is a good option. Truncated backpropagation works by splitting the long sequence into shorter sequences and treats each shorter sequence as a separate training case. Then the state of the parameters in the last time step of the current short sequence will be passed to the next sequence.

We conducted experiments by varying the number of time steps in the truncated backpropagation. The experiments were performed on the CCAT10 training set. We used 10% of the training data as the validation set. Table 4.5 presents the experimental results. As can be observed from the table, using a larger number of time steps results in higher average perplexity. This is more likely caused by the shortcoming of the truncated backpropagation. Although truncated backpropagation maintains the recurrent hidden state between the networks, however it truncates gradient flows between subsequences (Tallec and Ollivier, 2017). This causes the problem of learning dependencies above the range of truncation/time steps. Intuitively, a network with a larger number of time steps will be more affected. In addition, we argue that in the character-level language model, information from long context sequences is not useful for predicting the next character. However, based on our observation, training the LSTM language model with a smaller number of time steps required longer computation times as there is a greater number of training samples to be processed. In previous work (Zaremba et al., 2014; Kim et al., 2016) the number of time steps was set to fewer than 50.

We expect that more accurate language models will lead to better authorship attribution accuracy. However, the results in Table 4.5 demonstrate that the performance of authorship attribution is not really affected by the language model perplexity. The differences in authorship attribution accuracy for each number of time steps is more likely to be caused by the small size of the validation data. Our observation is that since the LSTM model is optimized for the language modeling task, perplexity is not a good predictor for the authorship attribution performance. More accurate language models are not guaranteed to improve the accuracy of authorship attribution. Previous studies utilizing language model for other NLP tasks reported similar conclusions (Zweig et al., 2012; Mirowski and Vlachos, 2015).

num_steps	AA accuracy	Average Perplexity
3	86.00	6.473473
10	86.00	6.804340
20	84.00	7.185383
30	88.00	7.479045
40	88.00	7.547656
50	88.00	7.684623
60	88.00	7.791094
200	88.00	8.218547

Table 4.5: Authorship attribution accuracy and average perplexity on the CCAT10 validation set

	n-gram LM		LSTM	
Dataset	AA acc	perplexity	AA acc	perplexity
Judgment	80.78	5.98	75.92	3.81
CCAT10	75.00	4.56	75.60	6.68
CCAT50	64.76	5.01	66.20	7.53
IMDb62	87.39	5.27	*86.03	5.45

Table 4.6: Authorship attribution accuracy and average perplexity on the test data using a character level n-gram and LSTM-based language model. *Due to the long training time of LSTM, for IMDb62 we performed the experiment only in the first fold (from 10-folds) of the dataset.

4.3.5 Limitation of LSTM-based language model for authorship attribution

Table 4.6 presents the performance of the LSTM-based model in four datasets. Compared to the n-gram language model-based approach, the LSTM-based model obtained slightly higher authorship attribution performance in both CCAT datasets but lower accuracy in two other datasets (Judgment and IMDb62). We argue that the small improvement obtained in CCAT10 and CCAT50 is due to the high topical diversity of the datasets which is better captured by the LSTM. These results confirm our previous analysis in the latter section (Section 4.2), that language model-based approaches are more suitable for datasets with more clear topical distinction between authors. Furthermore, we can observe from the table that perplexity is a bad indicator for authorship attribution performance.

The poor accuracy is more likely due to the use of models not optimized for the authorship attribution task. The LSTM results reported here also failed to outperform the previous work which used simpler approaches (see Table 4.7).

We observed that optimizing the LSTM-based language model separately from authorship attribution is ineffective since each author's language model needs to be optimized individually. This process involves hyper-parameters tuning and optimizing the model on the training data via backpropagation. Considering the authorship attribution dataset may consist of a large number of authors, the optimization process will be computationally expensive. A possible direction to address this problem is by jointly training the language model and authorship attribution via Multi Task Learning (MTL) (Caruana, 1993). In this way, the language modeling can be used as a second objective function for authorship attribution. Previous work applied this approach for sequence modeling tasks (Rei, 2017).

Model	Judgment	CCAT10	CCAT50	IMDb62
Previous work				
SVM with affix+punctuation 3-grams (Sapkota et al., 2015)	-	78.80	**69.30**	-
SVM with 2,500 most frequent 3-grams (Plakias and Stamatatos, 2008)	-	**80.80**	-	-
Token SVM (Seroussi et al., 2013)	91.15	-	-	**92.52**
Authorship attribution with topic models (Seroussi et al., 2013)	**93.64**	-	-	91.79
Character-level n-gram LM	80.78	75.00	64.76	87.39
Word-level n-gram LM	80.55	59.60	46.52	68.00
Character-level LSTM-based LM	75.92	75.60	66.20	86.03

Table 4.7: Comparison against previous results.

4.4 Summary

In this chapter we presented the evaluation of n-gram and LSTM-based language models for the authorship attribution task. We demonstrated how the n-gram language model can be used to address the feature selection problem which is commonly faced in the task. Furthermore, we implemented an LSTM-based language model which can capture information from longer context sequences. We provided a thorough analysis on the model performance and explored limitations of the approaches described.

Chapter 5

Continuous *N*-gram Representations for Authorship Attribution

In Chapter 4, it was demonstrated that information from long context sequences does not really benefit the authorship attribution performance. Our LSTM-based results failed to outperform the previous work which used simpler approaches such as linear classifier with bag-of-words features. Information from local structure which can be captured by character and word n-gram features, is likely to be more useful for authorship attribution (Kjell, 1994; Forsyth and Holmes, 1996; Stamatatos, 2006; Peng et al., 2003; Keselj et al., 2003; Juola, 2004; Stamatatos, 2013; Schwartz et al., 2013). Furthermore, word n-grams can represent document topic (Coyotl-Morales et al., 2006; Wang and Manning, 2012) while character n-grams have been shown to be effective for capturing stylistic, topical and morphological information (Koppel et al., 2011; Sapkota et al., 2015). However, previous work in authorship attribution mostly relied on discrete feature representations which suffer from sparsity and do not consider semantic relatedness between features (Mikros and Perifanos, 2013; Joulin et al., 2017).

This chapter aims to address this problem by introducing the use of continuous n-gram representations for authorship attribution tasks. Continuous representations have been shown to be helpful in a wide range of natural language processing tasks (Mikolov et al., 2013b; Bansal et al., 2014; Joulin et al., 2017; Li et al., 2016; Rahimi et al., 2017). Unlike previous work, each n-gram is represented in continuous vector space, and these representations are learned in the context of the authorship attribution tasks considered. More specifically, continuous n-gram representations are learned jointly with the classifier as a feed-forward neural network, combining the advantages of n-gram features and continuous representations. Furthermore, the model does not need any external linguistic resources such as Wordnet, that have limited coverage and/or are not available for many domains and languages. The proposed method outperforms the prior state-of-the-art approaches on two out of four datasets while producing comparable results on the remaining ones. In addition to that, we apply the results of the analysis in Chapter 3 via a novel extension of the proposed approach and obtain improvement on two datasets.

This chapter also explores the use of the author's demographic profiles. A significant amount of work focuses on demographic profile based tasks such as gender and age identification (see Section 2.3.3). Research in this area has found that there are differences in the writing of people from different demographic categories. Demographic profiles may provide information about the author of a document. Some previous work demonstrated that improvements can be obtained in some tasks by taking account of demographic information (Hovy, 2015; Benton et al., 2017). However there have been no previous attempts to explore this for the authorship attribution. In this chapter, Multi Task Learning (MTL) is used to jointly learn authorship attribution, gender and age. Experiments implement two different MTL models. Results show that incorporating gender and age information produces a small, but consistent improvement in performance.

The main content of this chapter is split into three sections. First, Section 5.1 presents the proposed continuous n-gram representation models and experiments which were performed on four different datasets (Judgment, CCAT10, CCAT50 and IMDb62). In this section, some results and analysis are also described. Second, in Section 5.2, we describe our novel extension of the proposed continuous n-gram representation models and show how the analysis in Chapter 3 helps to improve the attribution accuracy. Third, in Section 5.3 the proposed multi task learning models are presented. The discussion of the performed experiments on The Blog Authorship corpus is included in this section. Finally, the conclusions of this chapter are reported in the last section.

5.1 Continuous N-gram Representations

In this section, we present our proposed continuous n-gram representation models and describe our experiments in four authorship attribution datasets. We begin this section with a short overview of word embedding which is closely related to our proposed model.

Word embedding is a distributed word representation where an individual word is represented as a d-dimensional real-valued vector. Each word is mapped into a vector and the vector values are learned via a neural network-based model (Bengio et al., 2003; Mikolov et al., 2010). One of the main advantages of this representation is in generalization power (Goldberg and Hirst, 2017). Words which have similar semantic and grammatical roles are likely to have similar representations (Bengio et al., 2003). Given these following sentences: *The cat is walking in the bedroom* and *A dog was running in a room*, the word *cat* and *dog* will have similar representations as they are used in similar ways.

Mikolov et al. (2013c) proposed Word2Vec, a predictive model for learning a standalone word embedding from training corpus. He demonstrated that syntactic and semantic regularities in language can be captured using vector-space representations. The regularities simply can be characterized by a relation-specific vector offset. As an example, the distance between words *king* and *queen* are similar as the distance between words *man* and *woman* as illustrated in Figure 5.1. As part of Word2Vec, two different learning models are introduced: the Continuous Bag-of-Words model (CBOW) and the Skip-Gram model (Mikolov et al., 2013a). The CBOW model learns the word embedding by predicting the target words from source context words. In contrast, the Skip-Gram model predicts the surrounding words from the target words.

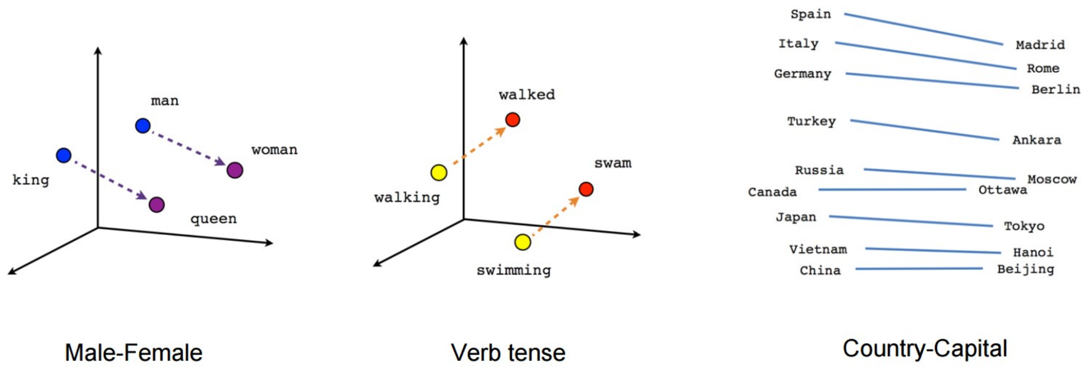


Figure 5.1: Illustration of word in vector space (source: https://www.tensorflow.org/)

We adopt the idea of word embedding in our proposed models. However, instead of learning word representations, our models learn the representations of n-gram features in the context of the authorship attribution task.

5.1.1 Model Architecture

A shallow neural network architecture, *fastText*, proposed by Joulin et al. (2017) was applied to learn the n-gram feature representations jointly with the classifier. This model is similar to a standard linear classifier, but instead of representing a document with a discrete feature vector, the model represents it with a continuous vector which is obtained by averaging the continuous vectors of the features present. More formally, *fastText* predicts the probability distribution over the labels for a document as follows:

$$\hat{y} = softmax(BAx) \tag{5.1}$$

where x is the bag of features for the document, the weight matrix A is a dictionary containing the embeddings learned for each feature, and B is a weight matrix that is learned to predict the label correctly using the resulting representations (essentially weighted feature embeddings). For a set of M documents, training involves minimizing the negative log-likelihood over the classes:

$$L = -\frac{1}{M}\sum_{m=1}^{M} y_m log\hat{y}_m \tag{5.2}$$

where y_m is the target distribution and $\hat{y}_m$ is the output distribution for a particular document m. The model is illustrated in Figure 5.2. The model consists of an embedding layer (hidden layer), average pool and output layer. The embedding layer is used to learn the continuous representations of the n-gram features. The representations are then averaged and fed into the output layer.

Since the documents in this model are represented as bags of discrete features, sequence information is lost. To recover some of this information feature n-grams are considered, similar to the way convolutional neural network architectures incorporate word order (Kim, 2014) but with a simpler architecture. Even though the proposed model ignores long-range dependencies in sentences that can be captured using recurrent neural network architectures that are commonly used in natural language processing tasks (Mikolov et al., 2010; Luong et al., 2013), topical or stylistic information mostly found in shorter word or character sequences for which the shallow neural network architecture with n-gram feature representations is likely to be sufficient, while much faster to run since the documents considered are much longer than the single sentences which RNNs typically model.

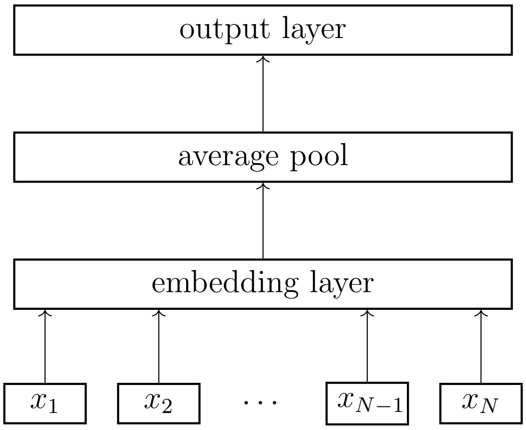


Figure 5.2: FastText model

5.1.2 Experiment

Experiments were performed using four datasets: Judgment, CCAT10, CCAT50 and IMDb62 (see Section 3.1).

5.1.2.1 Model Variations

Experiments were performed with three variations of the approach:

- **Continuous word *n*-grams**. In this model word unigrams and bigrams were used. The vocabulary size was set to 700 words.

- **Continuous character *n*-grams**. Following our previous experimental results in Section 4.1, n-grams up to and including a length of four were used. Previous work (Sanderson and Guenter, 2006), also found it to be the best n value for short English texts. Following Zhang et al. (2015) the vocabulary size was set to 70 characters including letters, digits, and some punctuation marks.

- **Continuous word and character *n*-grams**. This model combines word and character n-gram features.

5.1.2.2 Hyperparameters Tuning and Training Details

For all four datasets the Adam update rule (Kingma and Ba, 2014) was used to train the model. To avoid overfitting, validation loss was monitored using early stopping. Since none of the datasets have a standard development set, 10% of the training data was picked randomly for that purpose. Both word and character embeddings were initialized using Glorot uniform initialization (Glorot and Bengio, 2010). Following Keras's (Chollet, 2015) implementation of *fastText*, the softmax function was used in the output layer. This experiment did not use the *hashing trick* (Weinberger et al., 2009) which was unnecessary for relatively small sized datasets.

For the Judgment, CCAT10, and CCAT50 datasets, an embedding layer with a size of 100, dropout rate of 0.75, learning rate of 0.001 and mini-batch size of 5 were used. The number of epochs was set to 150. The values of dropout rate and mini-batch size were chosen via a grid search on the CCAT10 development set. Other hyper-parameter values (i.e. learning rate and embedding size) were fixed. For IMDb62, the same dropout rate as above was used. The learning rate, embedding size, mini-batch size and number of epochs were set to 0.01, 50, 32 and 20 respectively, by considering the dataset has a relatively large number of training instances.

5.1.3 Results and Discussion

Table 5.1 presents the comparison of the proposed approaches against the previous state-of-the-art methods on the four authorship attribution datasets considered. Overall, results show the effectiveness of continuous n-gram representations which outperform the previous best results on the CCAT50 and IMDb62 datasets. In the Judgment dataset, the models obtained comparable results with the previous best. However as can be seen in the table, the accuracy on CCAT10 is lower than the one reported in the previous work.

5.1.3.1 Word vs Character

The results in Table 5.1 demonstrate that performance is higher using character models. In particular, it is found that models which employ character level n-grams appear to be more suitable for datasets with a large number of authors, i.e. CCAT50 and IMDb62. To explore this further, an additional experiment was ran by varying the number of authors on a subset of IMDb62. For each of the authors 200 documents were used, with 10% of the data set as the development set and another 10% as the test set. Figure 5.3 shows a steep decrease in the accuracy of word models as the number of authors increases. The drop in accuracy of the character n-gram model is less pronounced.

Character models also achieved a slightly better result on the Judgment dataset which consists of only three authors. This can be explained by the fact that the documents in this corpus are significantly longer; almost ten and four times longer than those in IMDb62 and CCAT50 respectively (see Table 3.1).

Model	Judgment	CCAT10	CCAT50	IMDb62
Previous work				
SVM with affix+punctuation 3-grams (Sapkota et al., 2015)	-	78.80	69.30	-
SVM with 2,500 most frequent 3-grams (Plakias and Stamatatos, 2008)	-	**80.80**	-	-
Token SVM (Seroussi et al., 2013)	91.15	-	-	92.52
Authorship attribution with topic models (Seroussi et al., 2013)	**93.64**	-	-	91.79
Chapter 4				
Character-level n-gram LM	80.78	75.00	64.76	87.39
Word-level n-gram LM	80.55	59.60	46.52	68.00
Character-level LSTM-based LM	75.92	75.60	66.20	86.03
Proposed models				
Continuous n-gram words (1,2)	90.31	77.80	70.16	87.87
Continuous n-gram char (2,3,4)	91.29	74.80	**72.60**	**94.80**
Continuous n-gram words (1,2) and char (2,3,4)	91.51	77.20	72.04	94.28

Table 5.1: Comparison against previous results.

The large numbers of word n-grams make it more difficult for the model to learn good parameters for them. Combining word and character n-grams only produced a very small improvement on the Judgment dataset. We argue this was probably due to the more parameters to learn which are not suitable for the relatively small size datasets involved in our experiments. In CCAT10, character models were not as effective as word or character-word models. Our observation is that the high topical diversity between authors in CCAT10 is the main factor that influences the model's performance.

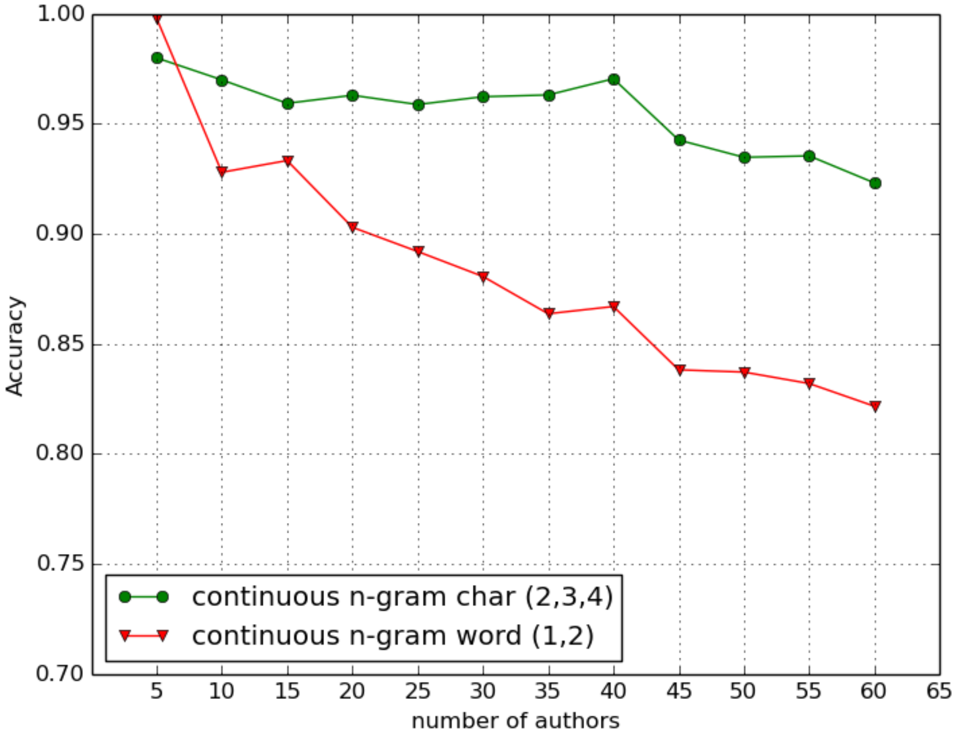


Figure 5.3: Accuracy on IMDb62 data subset with varying number of authors

5.1.3.2 Domain Influence

Most previous work on authorship attribution has concluded that content words are more effective for datasets where the documents can be discriminated by topic (Peng et al., 2004; Luyckx and Daelemans, 2010). Seroussi et al. (2013) show that the Judgment and IMDb62 datasets fall into this category and approaches based on topic models achieve high accuracy (more than 90%). However, the results in Table 5.1 demonstrate that stylistic information from continuous character n-grams outperforms word-based approaches on some datasets. In addition, these results also support the superiority of character n-grams that has been reported in previous work (Peng et al., 2003; Stamatatos, 2013; Schwartz et al., 2013). We observed that the low topical diversity between authors in Judgment and IMDb62 influences the effectiveness of word models in identifying the correct authors. On the other hand, character models provide information about the author's writing style which is more useful in this type of dataset. The experiments in Chapter 3 confirm these observations.

5.1.3.3 Feature Contributions

An ablation study was performed to further explore the influence of different types of features by removing a single class of n-gram features. For this experiment the character model was used. Three feature types are defined including:

1. **Punctuation *N*-gram:** A character n-gram which contains punctuation (e.g. `nts, ng" ars.  ay’s no-c`). There are 33 punctuation marks in total: ‘-’, ‘,’, ‘;’, ‘.’, ‘!’, ‘?’, ‘:’, ‘“’, “’, ‘\’,‘\\’, ‘—’, ‘_’, ‘@’, ‘#’, ‘$’, ‘%’, ‘~’, ‘&’,‘*’, ‘^’, “’, ‘+’, ‘=’, ‘<’, ‘>’, ‘(’, ‘)’, ‘[’, ‘]’, ‘{’, ‘}’, and the newline symbol.

2. **Space *N*-gram:** A character n-gram that contains at least one whitespace character (e.g. `to_ the_ by_a ng_`) .

3. **Digit *N*-gram:** A character n-gram that contains at least one digit (e.g. `F-16 6,7 3.1 PX50`) .

	Judgment (Δ)	CCAT10 (Δ)	CCAT50 (Δ)	IMDb62 (Δ)
all features (char model)	91.29	74.80	72.60	94.80
(–) punctuation n-grams	85.77 (-5.52)	73.80 (-1.00)	68.80 (-3.80)	87.90 (-6.90)
(–) space n-grams	85.55 (-5.74)	71.80 (-3.00)	70.20 (-2.40)	92.70 (-2.10)
(–) digit n-grams	90.91 (-0.38)	75.60 (+0.80)	71.28 (-1.32)	94.90 (+0.10)
(–) bi-grams	91.28 (-0.01)	76.20 (+1.40)	72.08 (-0.58)	95.12 (+0.32)
(–) tri-grams	86.88 (-4.41)	74.80 (0.00)	71.84 (-0.76)	95.40 (+0.60)
(–) four-grams	86.81 (-4.48)	74.40 (-0.40)	71.16 (-1.44)	92.76 (-2.04)

Table 5.2: Results of feature ablation experiment.

Table 5.2 demonstrates that removing punctuation, space and character four-grams leads to performance drops on all of the datasets. This is because some of those n-grams such as *the_ to_ and_* are function words which are essential on capturing the writing style of the authors. The author’s unconscious behavior in using punctuation is also a good feature for authorship attribution (Grieve, 2007; Sapkota et al., 2015). On the other hand, leaving out digit n-grams and bi-grams improves accuracy on the CCAT10 dataset. The CCAT10 dataset which was constructed from corporate/industrial on-line news contains some texts dominated by digits (e.g. articles related to the stock exchange). However, using digits in text usually needs to follow specific formating e.g. digits in time followed by AM/PM, a year is written in four digits format. Thus, these features are not really effective for identifying the author. Removing character tri-grams affects the performance on Judgment, since it may capture the author's writing styles which are important for this dataset. However, character tri-grams tend to be less useful on other datasets.

5.2 Extending The Continuous *N*-gram Representation Models

In Chapter 3, we have described experiments on the relationship between the effectiveness of different types of features for authorship attribution with different characteristics of datasets. We have presented feature ablation studies which covered three different types of features: *style, content* and *hybrid*. Content-based features tend to be suitable for datasets with high topical diversity such as the one constructed from on-line news. On the other hand, datasets with less topic variance e.g. legal judgment and movie review, fit with style-based features. In this section, we aim to further validate our findings in Chapter 3. We extend the model presented in Section 5.1 by incorporating each feature type (*style, content* and *hybrid*) as an auxiliary feature represented in discrete form. Auxiliary features provide additional information related to the dataset characteristics.

Given x_{aux} as a normalized auxiliary feature frequency vector, V is the weight applied to the features and f is the activation function (ReLu), the hidden layer h performs the following computation:

$$h = f(Vx_{aux}) \tag{5.3}$$

The probability distribution over the label for a document then can be described as:

$$\hat{y} = softmax(W_{out}[Ax, h]) \tag{5.4}$$

where x is the frequency vector of features for the document, A is the embedding matrix, W_{out} is the weight matrix of the output layer and $[Ax, h]$ is the concatenation vector of Ax and h. Figure 5.4 illustrates the model architecture.

For experiment in this section, we use the character-based model as the baseline, since it outperformed the state-of-the-art on the CCAT50 and IMDb62 datasets, while producing comparable results on the remaining two.

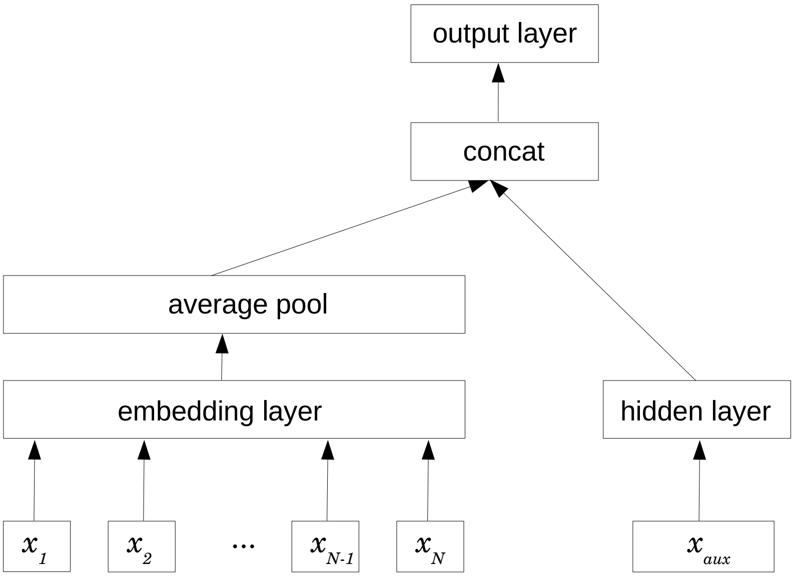


Figure 5.4: The extended continuous n-gram representation model with auxiliary features

5.2.1 Hyper-parameter Tuning

All character n-gram embeddings in the model were initialized using Glorot uniform initialization (Glorot and Bengio, 2010). We used the best hyper-parameter values for each of the datasets which have been tuned in the development set via a small grid search over all combinations of embedding size and dropout rate (specifically dropout in the concatenation layer). The size of the hidden auxiliary layer was set to 2. For the rest of the hyper-parameters, we used values from the baseline model (the continuous character n-grams). For Judgment, CCAT10 and CCAT50, we set the number of epochs to 250, and used 100 for IMDb62. For all datasets, early stopping was used on the development sets and the models were optimized with the Adam update rule (Kingma and Ba, 2014).

5.2.2 Experimental Results

Table 5.3 presents the results of the experiment. It can be seen that for each of the four data sets there is at least one feature type which leads to improved results when incorporated into the model. Our results demonstrate that better performance can be achieved by taking the data characteristics into account

Dataset	baseline	+style	+content	+hybrid
Judgment	91.29	91.07	**91.51**	91.21
CCAT10	74.80	76.00	**76.20**	74.80
CCAT50	72.60	72.72	**72.88**	71.76
IMDb62	94.80	**95.93**	95.59	95.26

Table 5.3: Extended model results

when choosing authorship attribution features. Moreover, the results provide evidence that character n-grams which have been known as the typical *go-to* features do not perform equally well in all types of datasets. For the three datasets (CCAT10, CCAT50 and IMDb62) the best results are obtained using the feature type identified as being most useful in Section 3.3. The only exception we found was that using the style features does not improve results on the Judgment dataset as we had expected. The relatively poor performance of the style features may be due to the baseline model (the continuous character n-grams) which effectively captured the author's writing style. Thus the addition of auxiliary style features did not lead to any improvement.

The results reported here for the CCAT50 and IMDb62 datasets outperform the previously best reported results presented in Section 5.1 and the model reported here therefore represents a new state-of-the-art performance (see Table 5.4).

Model	Judgment	CCAT10	CCAT50	IMDb62
Previous work				
SVM with affix+punctuation 3-grams (Sapkota et al., 2015)	-	78.80	69.30	-
SVM with 2,500 most frequent 3-grams (Plakias and Stamatatos, 2008)	-	**80.80**	-	-
Token SVM (Seroussi et al., 2013)	91.15	-	-	92.52
Authorship attribution with topic models (Seroussi et al., 2013)	**93.64**	-	-	91.79
Chapter 4				
Character-level n-gram LM	80.78	75.00	64.76	87.39
Word-level n-gram LM	80.55	59.60	46.52	68.00
Character-level LSTM-based LM	75.92	75.60	66.20	86.03
Chapter 5.1				
Continuous n-gram words (1,2)	90.31	77.80	70.16	87.87
Continuous n-gram char (2,3,4)	91.29	74.80	72.60	94.80
Continuous n-gram words (1,2) and char (2,3,4)	91.51	77.20	72.04	94.28
The extended model	91.51	76.20	**72.88**	**95.93**

Table 5.4: Comparison against previous results.

5.3 Improving attribution performance using the author's demographic profile

The task of authorship attribution can be seen as a way to model authors based on their characteristics such as topical preferences and writing style. A large number of studies have used similar characteristics to model the author's demographic profiles such as age, gender, personality and occupation (see Section 2.3.3). An interesting finding from recent work is that learning those demographic profiles jointly with text classification tasks might create opportunities to model coinciding influence factors among them (Benton et al., 2017).

As an example, Hovy (2015) evaluated the effect of age and gender information on classification performance in three NLP tasks: sentiment analysis, topic detection and author attribute classification. Results from his experiments show consistent improvements across tasks and languages. He argued that the differences between each demographic group's use of language are the main reason for improvement. Similar results were obtained by Benton et al. (2017) who used Multi Task Learning (MTL) to model multiple mental health conditions. By performing experiments with nine different auxiliary tasks, he observed that more accurate predictions were obtained when the right set of tasks are chosen. For example prediction of a *bipolar* condition achieved the best performance when prediction of suicide attempts and depression were used as auxiliary tasks. Furthermore, some tasks were also found to be similar: e.g. a model for predicting suicide attempts may also be good at predicting anxiety.

MTL has been known to help improve the performance of single task models (STL) (Caruana, 1993). Caruana argued that information provided by auxiliary tasks act as a domain-specific inductive bias for the main tasks. However, previous work reported mixed results which shows MTL does not always guarantee improvement (Klerke et al., 2016; Luong et al., 2016; Martínez Alonso and Plank, 2017; Søgaard and Goldberg, 2016). Studies found that the performance improvement is heavily influenced by the choice of auxiliary tasks (Bingel and Søgaard, 2017).

In this section, experiments were conducted by extending the first model in Figure 5.2 to use MTL. Using an MTL framework, the authorship attribution tasks are learned in parallel with two other tasks: age and gender identification. According to Schler et al. (2006), there are noteworthy differences between demographic groups' use of certain stylistic and content features. The most notable differences are in the usage of blog words, pronouns, determiners and prepositions. Compared to other age groups, teenagers more commonly use blog words such as *lol, haha, ur* and pronouns such as *I, you, she, he.* In addition, their averaged post lengths are the shortest among the age groups. In contrast, people in the 30s age group tend to use more prepositions and determiners in their writings. Different writing styles are also found between gender groups. Pronouns, negation words and blog words are more commonly used by females. On the other hand, male bloggers use more hyperlinks. In terms of content, topical preferences of each age group reflect their concerns at the time of writing. For example, teenagers are more concerned about friends and mood swings. Thus, their posts are dominated by topics related to *happy, boring, homework.* People in their 20s discuss more about college life, since most of them are students, while topics related to marriage, family life, financial concerns and politics are commonly discussed by people in their 30s. People in different gender groups are also have different topical interests. Female bloggers tend to write more 'personal' writing, while posts by male bloggers are dominated by topics related to politics and technology (see A.2 for statistic of The Blog Authorship Corpus). This section will explore the benefits of age and gender identification for authorship attribution.

5.3.1 Multi Task Learning (MTL) model

In this experiment, we used hard parameter sharing in a deep neural network approach (Caruana, 1993). This approach works by sharing some of the hidden layers so that it allows the model to learn a joint representation for multiple tasks. We used the character-level neural network model presented in Section 5.1 as the baseline architecture for our MTL experiments. Our MTL model is presented in Figure 5.5. First, an averaged continuous representation for a document is learned via an embedding layer (hidden layer). This joint-tasks representation is then fed into the corresponding task-specific output layers. The prediction probabilities for a particular task p is computed by modifying the Equation 5.1:

$$\hat{y}^{(p)} = softmax(B^{(p)}Ax) \tag{5.5}$$

where $B^{(p)}$ is the weight matrix in the output layer. Given P related tasks, the global loss function is the linear combination of the loss function for all tasks.

$$\phi = \sum_{p=1}^{P} \lambda_p L(\hat{y}^{(p)}, y^{(p)}) \tag{5.6}$$

$L(\hat{y}^{(p)}, y^{(p)})$ is loss function for particular task p (see Equation 5.2) and λ_p is the weights for each task p which is used to control the importance of the task's loss.

The hard parameter sharing model is easy to implement and known to be an efficient regularizer (Baxter, 2000; Søgaard and Goldberg, 2016)

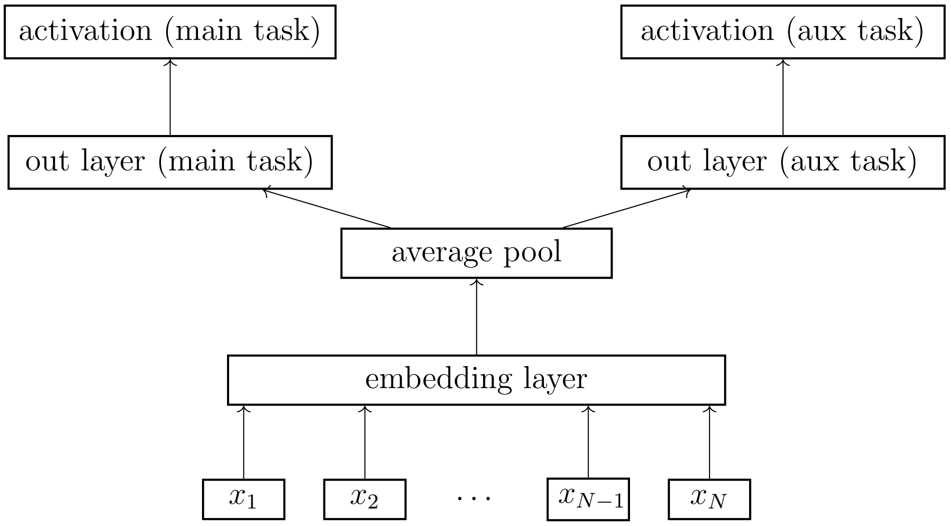


Figure 5.5: Multi task learning (MTL) model

5.3.2 Experiments

Experiments were conducted to observe the effect of auxiliary tasks (gender and age identification) on the main task (authorship attribution).

5.3.2.1 Data set

For this experiment, subsets of The Blog Authorship Corpus (Schler et al., 2006) were used. The corpus consists of 681,288 posts and over 140 million words from 19,320 bloggers collected from blogger.com in August 2004. Unlike the other datasets (Judgment, CCAT10, CCAT50 and IMdb62), this corpus provides additional information about the author including gender, age, industry and astrological sign. The Blog Authorship corpus consists of three age groups: 10s (age 13-17) with 8240 posts, 20s (age 23-27) with 8086 posts and 30s (age 33-47) with 2994 posts. For each age group there is an equal number of male and female bloggers. Although this corpus contains a large number of authors, the number of posts per author varies. In addition to that, as the blog posts are considered more informal, some authors can have either very short posts (consisting of only a few words) or relatively long posts (see Figure 5.6). The corpus is available for download from the following link `http://u.cs.biu.ac.il/~koppel/BlogCorpus.htm`

SAMPLE POST 1

distorted electric guitars Duh, duh duhn ... da, da, da, da, dah ... duhn, dunh! I am 45% evil? I could go either way. I have sinned quite a bit but I still have a bit of room for error. My life is a tug of war between good and evil. Are you evil find out at Hilowitz.com Silly on-line tests!!! I demand a re-count! Some of the questions were only true because of stuff from my past, so I'm probaby less evil now, maybe 35% evil or so. Either way, I'm more good that evil . . . so the battle is not lost. Must . . . keep . . . fighting . . . against . . . evil!

SAMPLE POST 2

okay im having trouble the blog so bear with me......

Figure 5.6: Snippets from The Blog Authorship Corpus

For the purpose of the experiments in this section, only authors with at least 800 posts were selected and 500 posts with a minimum of 200 characters were chosen for each of the authors. The numbers for each gender and age group were balanced.

5.3.2.2 Model

Four different models were implemented, including non-neural network Single Task Learning (STL) models as the baseline approaches. Details for each of the models are given as follows:

- **Single Task Learning Feed- Forward Neural Network Model (STL-FNN)**
  For the single task learning model, experiments were performed by using the character-level feed forward neural model described previously in Section 5.1. The model works by representing character n-gram features by continuous representations and learning the representations jointly with the authorship attribution classifier.

- **Multi Task Learning Model (MTL1)**
  In this model, authorship attribution was jointly learned with two auxiliary tasks: age and gender identification. Figure 5.5 illustrates the multi task model where the embedding layer is shared between all tasks. Each task then has its own output layer and activation functions. The model is trained to minimize the global loss function (Eq. 5.6). Shared weights (A) and all task-specific weights ($B^{(p)}$) are updated in parallel.

- **Multi Task Learning Model (MTL2)**: This model is similar to MTL1 with a slightly different procedure on the training phase. First, both shared weights (A) and all tasks-specific weights ($B^{(p)}$) were trained jointly for n-epochs. After this initial joint training, only the authorship attribution-specific weights (main task) were updated for another m-epochs. Using this approach, previous work by Benton et al. (2017) reported significant improvement over the first model (MTL1). They argued that updating shared weights (A) and all task-specific weights ($B^{(p)}$) in parallel, might cause lower performance on different tasks, even though the global loss is decreasing. This can be minimized by training only the main task weights for more epochs.

- **Non-Neural Network Single Task Learning (STL non-NN)** For the non-neural network STL, we used a Support Vector Machine (SVM) and Logistic Regression (LR). Similar to the STL-FNN and MTL models, for both SVM and LR, we used character bi-grams to four-grams as features. For each n-gram we used a total of 2500 features. This is an optimal value according to our experimental results in Section 4.1.3 and previous work (Stamatatos, 2013). Results on the validation set show the RBF kernel performs better than the linear kernel for SVM. All LR hyper-parameters were set to default. We used SVM and LR implementations from Scikit Learn (Pedregosa et al., 2011).

model	aux task	batch size	learning rate
STL-FNN	-	(5,5,5,5,5,5)	(5x10^{-4};5x10^{-4};5x10^{-4};5x10^{-4};5x10^{-4};5x10^{-4})
MTL1	age	(10,5,10,25,25,25)	(5x10^{-4};10^{-3};10^{-3};5x10^{-4};5x10^{-4};5x10^{-4})
	gender	(10,25,5,5,5,10)	(5x10^{-4};5x10^{-4};10^{-3};5x10^{-4};5x10^{-4};10^{-3})
	age & gender	(10,10,25,5,5,25)	(5x10^{-4};5x10^{-4};10^{-3};5x10^{-4};10^{-3};5x10^{-4})
MTL2	age	(5,10,10,10,25,25)	(5x10^{-4};10^{-3};10^{-3};5x10^{-4};5x10^{-4};5x10^{-4})
	gender	(5,25,5,10,25,25)	(5x10^{-4};5x10^{-4};10^{-3};5x10^{-4};5x10^{-4};10^{-3})
	age & gender	(25,10,10,25,25,25)	(5x10^{-4};5x10^{-4};10^{-3};5x10^{-4};10^{-3};5x10^{-4})

Table 5.5: Details of optimum hyper-parameters for each model and each number of authors

5.3.2.3 Hyper-parameter tuning and training details

In order to provide a fair comparison, certain hyper-parameters were set to the same values for both STL and MTL including embedding size (150), layer initialization (Glorot Uniform) and dropout (0.75). Due to limited computational resources, only two hyper-parameters were optimised, batch size (5, 10, 25) and learning rate (0.001, 0.0005). The process for searching optimum hyper-parameters was done via grid search. Adam was chosen as the default optimizer, as it converges relatively faster compared to the other optimizers. The number of epochs was set to 200. Validation losses during training were monitored using early stopping. Details of the optimum hyper-parameters for each model are presented in Table 5.5. The single task model (STL-FNN) tends to produce the best performances with small batch sizes and learning rates. Multi task models achieved their optimum performances with larger batch sizes.

5.3.2.4 Results and Analysis

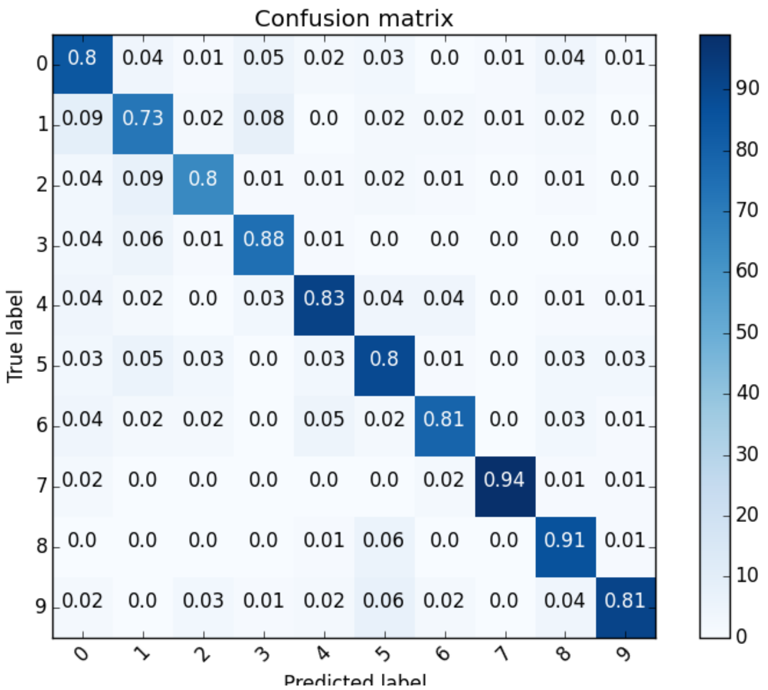
n_authors	STL non-NN		STL-FNN	MTL1			MTL2		
	SVM	LR		(a)	(g)	(a+g)	(a)	(g)	(a+g)
2	95.10	95.40	**95.50**	95.40	95.40	95.35	95.45	95.40	95.25
4	89.23	**91.30**	89.58	89.90	89.98	89.98	89.625	90.00	89.88
10	74.52	75.96	78.77*	78.73	78.51	**78.98***†	78.58	78.38	78.74
20	63.67	64.63	71.28*	71.35	71.34	71.29	71.27	71.25	**71.38***
30	58.12	59.28	66.64*	66.87	66.79	66.82	66.87	66.91	**66.93***
40	52.31	52.97	62.36*	62.25	62.39	**62.41***	62.29	62.36	62.37

Table 5.6: Accuracy of several models on the authorship attribution task. MTL results were obtained with certain auxiliary tasks including: (a):age identification; (g):gender identification; (a+g):age and gender identification. Significant improvement over the LR baseline at p=0.05 is denoted by * and over the STL model by †

Results for all models are presented in Table 5.6. We performed experiments with different numbers of authors. For each number of authors, results were averaged across 10 different sets. It is clear that neural network models (STL-FNN, MTL1 and MTL2) consistently produced better accuracies compared to non-neural network models (SVM and LR) particularly on the tasks with higher numbers of authors. Starting from 10 authors, there is a significant difference in the accuracy, while for fewer authors (2 and 4), non-neural models obtained similar or even better results. The results imply that the size of training data becomes an important factor which influences the neural network models performance. The size of training data grows as the more numbers of authors are added.

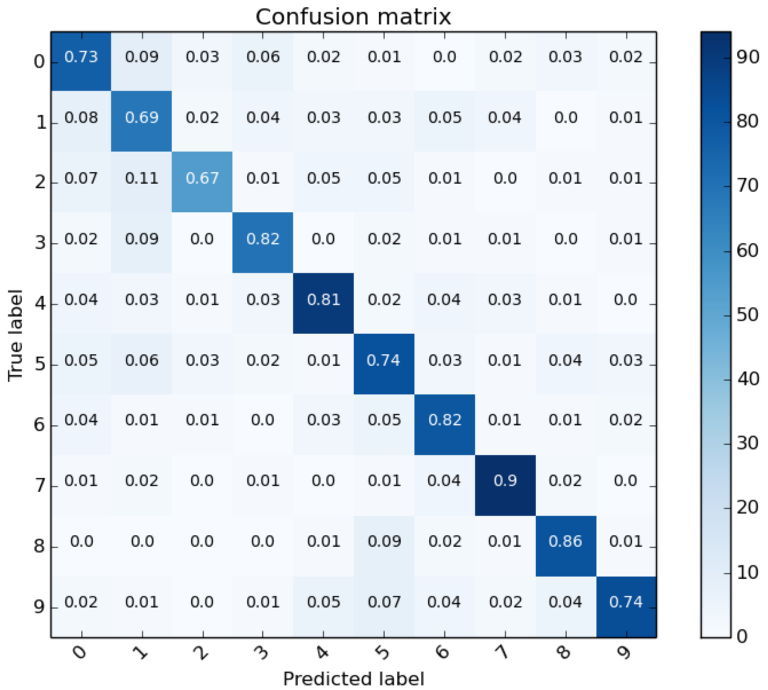
To gain insights into the classification results, confusion matrices were generated for each of the models. Figures 5.7, 5.8, 5.9 and 5.10 show confusion matrices for authorship attribution with 10 authors. The rows indicate the predicted label, whereas the correct label is indicated by the columns. For experiments with 10 authors, the best MTL1 result was obtained when using the model with age and gender as the auxiliary tasks. From the matrices, it can be seen that MTL1 successfully outperformed the non-neural network STL (SVM and LR) and STL-FNN models (see Figure 5.7). As can be observed from Figures 5.8, 5.9 and 5.10, all the STL models (SVM, LR and STL-FNN) made more incorrect predictions which mistakenly labeled author 7 as one of the other authors. MTL1 with age and gender identification improved the performance up to 94% in accuracy by lowering the incorrect predictions in the other classes (0, 1, 3, 4). We argue that the improvement is due to the variations of language use between each demographic group. As an example, authors 1 and 7 are identified as males from different age groups (author 1 in his 20s, while author 7 in the 10s) and with different topical interests. Blog posts written by author 1 are dominated by topics related to *computer, internet, email, and server*, while author 7 mostly discussed topics related to *anime, manga, japanese, and movie.*

The effectiveness of MTL models is also supported by the results from Table 5.6 which indicate consistent improvements with an increasing number of authors. This results agree with the previous experiments conducted by Hovy (2015) and Benton et al. (2017). Age and gender identification serve as regularizers for authorship attribution. As shown in Figures 5.11 and 5.12, training and validation losses in MTL1 converged more slowly than STL-FNN. However, it is hard to identify which auxiliary task has the most important contribution towards the improvements. As can be seen from Table 5.6, in most cases adding both age and gender identification result in higher accuracy gains. Restricting



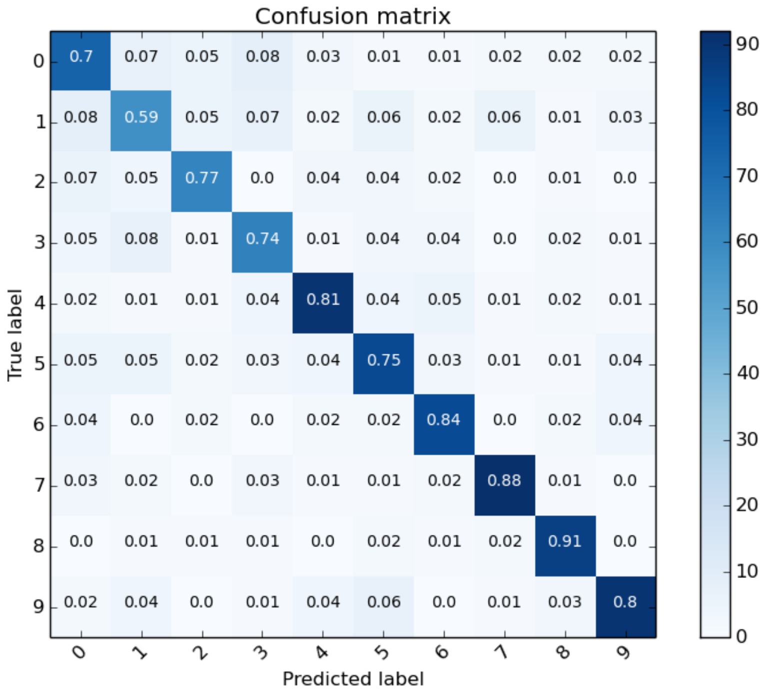
	0	1	2	3	4	5	6	7	8	9
0	0.8	0.04	0.01	0.05	0.02	0.03	0.0	0.01	0.04	0.01
1	0.09	0.73	0.02	0.08	0.0	0.02	0.02	0.01	0.02	0.0
2	0.04	0.09	0.8	0.01	0.01	0.02	0.01	0.0	0.01	0.0
3	0.04	0.06	0.01	0.88	0.01	0.0	0.0	0.0	0.0	0.0
4	0.04	0.02	0.0	0.03	0.83	0.04	0.04	0.0	0.01	0.01
5	0.03	0.05	0.03	0.0	0.03	0.8	0.01	0.0	0.03	0.03
6	0.04	0.02	0.02	0.0	0.05	0.02	0.81	0.0	0.03	0.01
7	0.02	0.0	0.0	0.0	0.0	0.0	0.02	0.94	0.01	0.01
8	0.0	0.0	0.0	0.0	0.01	0.06	0.0	0.0	0.91	0.01
9	0.02	0.0	0.03	0.01	0.02	0.06	0.02	0.0	0.04	0.81

Figure 5.7: Confusion matrix of MTL1 on the authorship attribution task with 10 authors



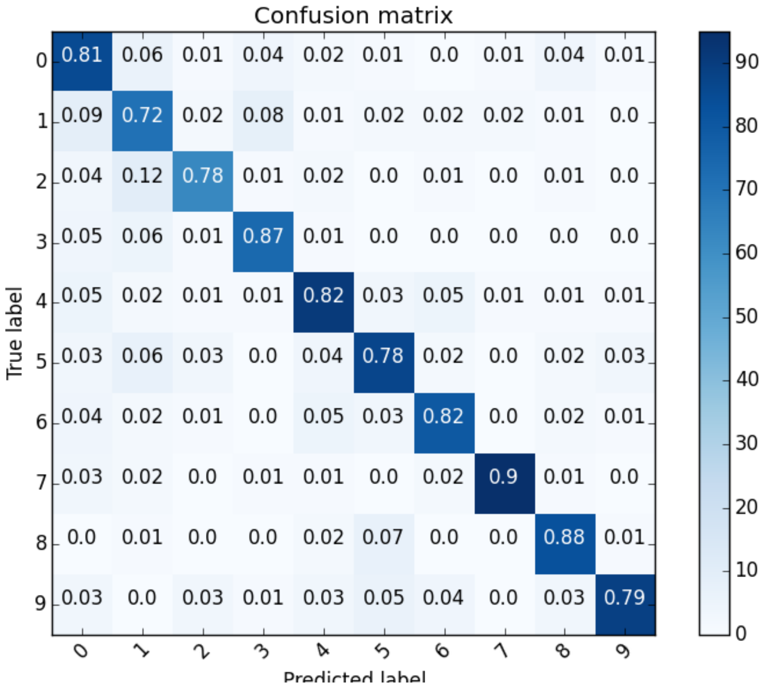
	0	1	2	3	4	5	6	7	8	9
0	0.73	0.09	0.03	0.06	0.02	0.01	0.0	0.02	0.03	0.02
1	0.08	0.69	0.02	0.04	0.03	0.03	0.05	0.04	0.0	0.01
2	0.07	0.11	0.67	0.01	0.05	0.05	0.01	0.0	0.01	0.01
3	0.02	0.09	0.0	0.82	0.0	0.02	0.01	0.01	0.0	0.01
4	0.04	0.03	0.01	0.03	0.81	0.02	0.04	0.03	0.01	0.0
5	0.05	0.06	0.03	0.02	0.01	0.74	0.03	0.01	0.04	0.03
6	0.04	0.01	0.01	0.0	0.03	0.05	0.82	0.01	0.01	0.02
7	0.01	0.02	0.0	0.01	0.0	0.01	0.04	0.9	0.02	0.0
8	0.0	0.0	0.0	0.0	0.01	0.09	0.02	0.01	0.86	0.01
9	0.02	0.01	0.0	0.01	0.05	0.07	0.04	0.02	0.04	0.74

Figure 5.8: Confusion matrix of SVM on the authorship attribution task with 10 authors



	0	1	2	3	4	5	6	7	8	9
0	0.7	0.07	0.05	0.08	0.03	0.01	0.01	0.02	0.02	0.02
1	0.08	0.59	0.05	0.07	0.02	0.06	0.02	0.06	0.01	0.03
2	0.07	0.05	0.77	0.0	0.04	0.04	0.02	0.0	0.01	0.0
3	0.05	0.08	0.01	0.74	0.01	0.04	0.04	0.0	0.02	0.01
4	0.02	0.01	0.01	0.04	0.81	0.04	0.05	0.01	0.02	0.01
5	0.05	0.05	0.02	0.03	0.04	0.75	0.03	0.01	0.01	0.04
6	0.04	0.0	0.02	0.0	0.02	0.02	0.84	0.0	0.02	0.04
7	0.03	0.02	0.0	0.03	0.01	0.01	0.02	0.88	0.01	0.0
8	0.0	0.01	0.01	0.01	0.0	0.02	0.01	0.02	0.91	0.0
9	0.02	0.04	0.0	0.01	0.04	0.06	0.0	0.01	0.03	0.8

Figure 5.9: Confusion matrix of LR on the authorship attribution task with 10 authors



	0	1	2	3	4	5	6	7	8	9
0	0.81	0.06	0.01	0.04	0.02	0.01	0.0	0.01	0.04	0.01
1	0.09	0.72	0.02	0.08	0.01	0.02	0.02	0.02	0.01	0.0
2	0.04	0.12	0.78	0.01	0.02	0.0	0.01	0.0	0.01	0.0
3	0.05	0.06	0.01	0.87	0.01	0.0	0.0	0.0	0.0	0.0
4	0.05	0.02	0.01	0.01	0.82	0.03	0.05	0.01	0.01	0.01
5	0.03	0.06	0.03	0.0	0.04	0.78	0.02	0.0	0.02	0.03
6	0.04	0.02	0.01	0.0	0.05	0.03	0.82	0.0	0.02	0.01
7	0.03	0.02	0.0	0.01	0.01	0.0	0.02	0.9	0.01	0.0
8	0.0	0.01	0.0	0.0	0.02	0.07	0.0	0.0	0.88	0.01
9	0.03	0.0	0.03	0.01	0.03	0.05	0.04	0.0	0.03	0.79

Figure 5.10: Confusion matrix of STL-FNN on the authorship attribution task with 10 authors

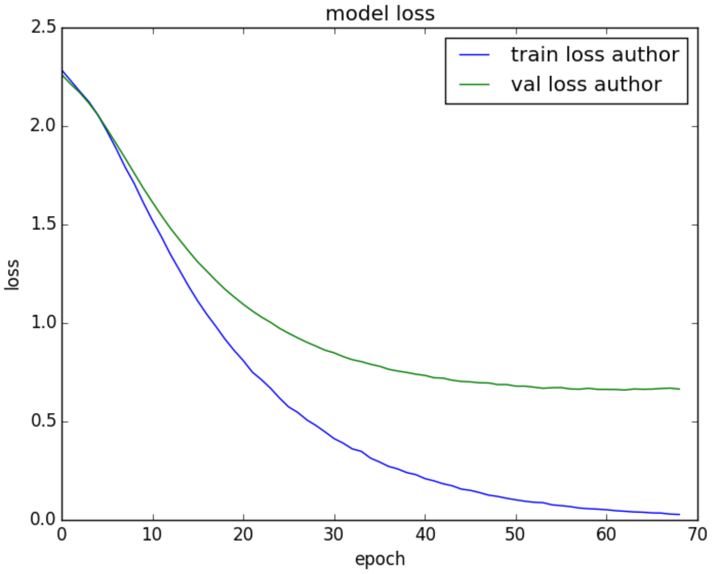


Figure 5.11: Training and validation losses of STL-FNN models

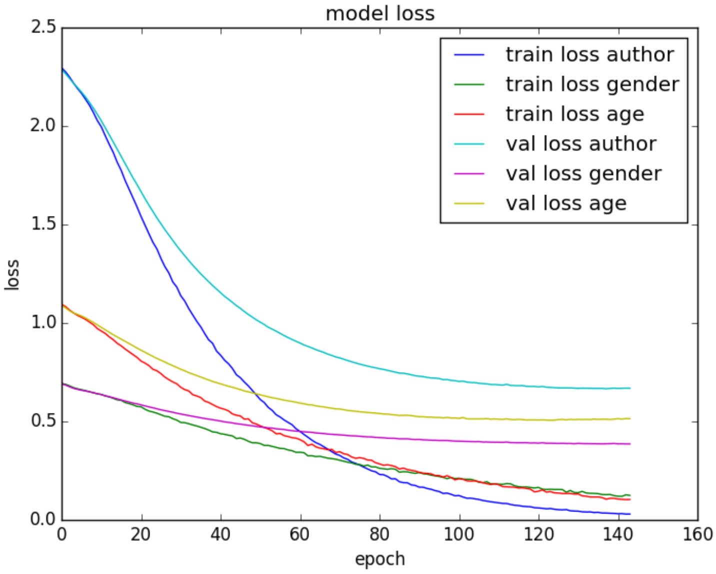


Figure 5.12: Training and validation losses of MTL1 models for authorship attribution with 10 authors (MTL1 model with the best accuracy).

the auxiliary task tends to hurt the accuracy of the attribution task. In contrast to Benton et al. (2017), we observed that there is no significant improvement obtained by training only the main task weights for more epochs (MTL2).

Nevertheless, apart from the obtained improvement, we found that training the MTL framework is challenging. Through the performed experiments, we observed that MTL is sensitive to hyper-parameter settings (e.g. batch size and learning rate). Furthermore, as has been pointed out in previous work (Caruana, 1996; Bingel and Søgaard, 2017; Benton et al., 2017) the choice of auxiliary task is essential in regard to the main task performance.

5.4 Summary

This chapter proposed continuous n-gram representations for authorship attribution. Using four authorship attribution datasets, it has been shown that the proposed model is accurate in identifying the writing style of the authors when compared to a strong baseline. Further improvement was obtained by incorporating auxiliary feature types discussed in Chapter 3 into the proposed model. In addition to that, experiments using MTL demonstrate the benefit of age and gender identification to the authorship attribution task. Results in experiments performed on The Blog Authorship corpus demonstrate the effectiveness of both tasks as a regularizer for authorship attribution as MTL yields small but consistent improvement over single task learning.

Chapter 6

Conclusions and Future Work

This thesis explored a variety of authorship attribution approaches in different types of datasets. This chapter presents a summary of contributions and findings throughout the thesis and proposes directions for future research.

6.1 Summary of Thesis Contributions

As stated in Chapter 1, the main goal of this thesis was to provide more clear direction of the authorship attribution approaches by exploring four different techniques implemented on various types of datasets. We believe that we have achieved this goal by tackling five subproblems: (1) the difficulties in determining which types of information will be most useful for a particular authorship attribution dataset; (2) evaluation of the effectiveness of n-gram language models on various types of datasets; (3) the limitation of n-gram-based language model on capturing information from long context sequences; (4) data sparsity and the inability of discrete representations to capture semantic relatedness between features; and (5) implementation of an authors' demographic profiles to improve authorship attribution performance.

In Chapter 3, we proposed a novel analysis using topic modeling to examine the conditions under which each type of authorship attribution feature is useful. Results from the analysis showed style-based features are more effective for datasets in which authors discuss similar topics. On the other hand, content-based features are generally more effective when there is more diversity between the topics discussed in the dataset. The hybrid features appear to behave similarly to the content-based features since they are most useful when the topic diversity is high. In addition, we have demonstrated that the results of the analysis are useful for both neural and non-neural network authorship attribution models.

In Chapter 4, we focused on investigating the effectiveness of language models for authorship attribution. We begin our contributions by presenting the feature selection problems that usually occur in authorship attribution. We have proved that the optimal values in feature selection which are suggested by the previous work are not applicable for all types of datasets. We argue that factors such as an imbalanced dataset and topical similarity between authors determine the optimal values in feature selection. This analysis prompted us to implement the n-gram language model approach for authorship attribution which has been claimed in previous work (Peng et al., 2003) to be effective for tackling the feature selection problem. Using a variety of datasets, we have demonstrated that the approach is more suitable for a dataset with higher topical diversity and/or distinct idionsyncracies in writing style. Our final contribution in this area is that we developed an LSTM-based language model for authorship attribution. From the experimental results, we argue that information from long context sequences does not benefit authorship attribution. Furthermore, we found that the language model's perplexity is not a good indicator for authorship attribution accuracy. Overall, throughout this chapter we have identified the strengths and limitations of the language model-based method.

In Chapter 5, we proposed a continuous n-gram representation for authorship attribution. Our model learns continuous representations for n-gram features via a neural network jointly with the classification layer. The representations addressed the problem in discrete feature representations which suffer from data sparsity and do not consider the semantic relatedness between features. Unlike previous work, each n-gram is represented in continuous vector space, and these representations are learned in the context of the authorship attribution tasks considered. The proposed model outperforms the state-of-the-art on two datasets, while producing comparable results on the remaining two. In addition, we describe our novel extension of the proposed models and show how the analysis in Chapter 3 helps to improve the attribution accuracy. Another contribution is that we proposed a Multi Task Learning (MTL) model which jointly learned an authorship attribution task with gender and age identifications. Results of the experiments demonstrate the effectiveness of both tasks as regularizers for authorship attribution. MTL yields small but consistent improvement over single task learning.

In sum, in this thesis we have validated four authorship attribution models on various types of datasets. We have demonstrated that in different circumstances, each model and feature representation may achieve different levels of accuracy.

Furthermore, we have provided suggestions on models and features which can produce optimal performance given a certain type of dataset.

6.2 Future Directions

A range of authorship attribution approaches explored in this thesis can be extended into different domains or attribution forms. We outline directions for future work:

- **Constructing Standard Authorship Attribution Datasets**
  An important area of future work is to construct standard datasets which enable evaluation of different methods of authorship attribution. We argue that researchers in authorship attribution are limited in exploring more advanced methods such as neural networks due to the unavailability of datasets with large numbers of training sets. Nevertheless, constructing an authorship attribution datasets is a challenging task. Juola (2008) and (Rudman, 2012) emphasised that a good dataset has to be constructed from clean and original writings by the authors. We believe standard datasets will encourage the continuity of systematic work in the authorship attribution field.

- **Exploring Different Languages**
  The majority of previous authorship attribution work has focused on English documents. There are only a few experiments which have been performed on languages other than English (Peng et al., 2003; Mikros and Perifanos, 2013). The PAN evaluation forum organized shared tasks in several languages, such as Dutch, Greek and Spanish (see Section 2.2), but research progress for these languages are relatively slow. A possible direction to explore is developing an adaptive approach which can be used for any language.

- **Multi Task Learning for Authorship Attribution**
  In Section 4.3.5 we identified the limitations of LSTM-based language model for authorship attribution. The model can be extended by jointly learning the language model and authorship attribution task via Multi Task Learning (MTL). Rei (2017) recently proposed a semi-supervised MTL for sequence labeling. It used the language model as the secondary training objective for several sequence labeling tasks such as error detection in text, named entity recognition, chunking and POS-tagging. The same method

can be implemented for authorship attribution, so that the task can be optimized on both the language model and the authorship attribution objective functions.

- **Neural Style Transfer for Authorship Obfuscation**
  Another possible extension of this work is to explore tasks related to authorship attribution such as authorship obfuscation. Section 2.3.5 provides a short review of this task. In authorship obfuscation, the original texts are modified so that the true author can not be identified. One possible method is by implementing an encoder-decoder LSTM (Bakhteev and Khazov, 2017). The encoder reads the the original texts and represents them as fixed-length embedding vectors. The decoder then decodes the vector and produces the obfuscated texts. Another potential approach is neural style transfer (Gatys et al., 2015) which was originally applied for images. The method works by separating and recombining the content and style of arbitrary images. A similar idea can be adopted for text documents. The content and style of original documents are recombined with other author's writings to produce obfuscated texts.

Bibliography

Abbasi, A. and Chen, H. (2008). Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)*, 26(2):1–29.

Afroz, S., Brennan, M., and Greenstadt, R. (2012). Detecting Hoaxes, Frauds, and Deception in Writing Style Online. In *Proceeding of the IEEE Symposium on Security and Privacy*, pages 461–475.

Argamon, S. and Levitan, S. (2005). Measuring the usefulness of function words for authorship attribution. In *Proceedings of the 2005 ACH/ALLC Conference*, pages 4–7.

Argamon-Engelson, S., Koppel, M., and Avneri, G. (1998). Style-based text categorization: What newspaper am I reading? In *Proceedings of AAAI Workshop on Learning for Text Categorization*, pages 1–4.

Baayen, H., van Halteren, H., and Tweedie, F. (1996). Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3):121–132.

Bagnall, D. (2015). Author Identification using multi-headed Recurrent Neural Networks—Notebook for PAN at CLEF 2015. In Cappellato, L., Ferro, N., Jones, G., and San Juan, E., editors, *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France.* CEUR-WS.org.

Bakhteev, O. and Khazov, A. (2017). Author Masking using Sequence-to-Sequence Models—Notebook for PAN at CLEF 2017. In Cappellato, L., Ferro, N., Goeuriot, L., and Mandl, T., editors, *CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers, 11-14 September, Dublin, Ireland.* CEUR-WS.org.

Bansal, M., Gimpel, K., and Livescu, K. (2014). Tailoring continuous word representations for dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 809–815, Baltimore, Maryland. Association for Computational Linguistics.

Baxter, J. (2000). A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12(1):149–198.

Benedetto, D., Caglioti, E., and Loreto, V. (2002). Language trees and zipping. *Physical Review Letters*, 88(4):48702.

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166.

Benton, A., Mitchell, M., and Hovy, D. (2017). Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 152–162, Valencia, Spain. Association for Computational Linguistics.

Bingel, J. and Søgaard, A. (2017). Identifying beneficial task relations for multitask learning in deep neural networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 164–169, Valencia, Spain. Association for Computational Linguistics.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Bogdanova, D. and Lazaridou, A. (2014). Cross-Language Authorship Attribution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, number 1, pages 26–31, Reykjavik, Iceland. European Language Resources Association (ELRA).

Brennan, M. (2012). *Managing quality, identity and adversaries in public discourse with machine learning*. PhD thesis, Philadelphia, PA, USA.

Brennan, M., Afroz, S., and Greenstadt, R. (2012). Circumventing Authorship Recognition to Preserve Privacy and Anonymity. *ACM Transactions on Information and System Security*, 15(3):1–22.

Brennan, M. and Greenstadt, R. (2009). Practical Attacks Against Authorship Recognition Techniques. In *Proceeding of the Twenty-First Innovative Applications of Artificial Intelligence Conference*, pages 60–65, Pasadena, California. AAAI.

Brinegar, C. S. (1963). Mark Twain and the Quintus Curtius Snodgrass Letters: A Statistical Test of Authorship. *Journal of the American Statistical Association*, 58(301):85–96.

Burrows, J. (1987). Word Patterns and Story Shapes: The Statistical Analysis of Narrative Style. *Literary and Linguistic Computing*, 2:61–70.

Burrows, J. (1992). Not Unles You Ask Nicely: The Interpretative Nexus Between Analysis and Information. *Literary and Linguistic Computing*, 7(2):91–109.

Burrows, J. (2002). 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3):267–287.

Cai, D., He, X., rong Wen, J., Han, J., ying Ma, W., Cai, D., He, X., rong Wen, J., Han, J., and ying Ma, W. (2006). Support Tensor Machines for Text Categorization. Technical report, University of Illinois as Urbana-Champaign.

Caliskan-Islam, A., Harang, R., Liu, A., Narayanan, A., Voss, C., Yamaguchi, F., and Greenstadt, R. (2015). De-anonymizing programmers via code stylometry. In *Proceedings of the 24th USENIX Conference on Security Symposium*, SEC'15, pages 255–270, Berkeley, CA, USA. USENIX Association.

Caruana, R. (1993). Multitask learning: A knowledge-based source of inductive bias. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 41–48, Amherst, MA, USA. Morgan Kaufmann.

Caruana, R. (1996). Algorithms and applications for multitask learning. In *In Proceedings of the Thirteenth International Conference on Machine Learning*, pages 87–95. Morgan Kaufmann.

Chaski, C. E. (2005). Who ' s At The Keyboard ? Authorship Attribution in Digital Evidence Investigations. *International Journal of Digital Evidence*, 4(1):1–13.

Chen, B. (2014). Introduction to srilm toolkit.

Chen, S. F. and Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, ACL '96, pages 310–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cheng, N., Chandramouli, R., and Subbalakshmi, K. (2011). Author gender identification from text. *Digital Investigation*, 8(1):78–88.

Chollet, F. (2015). Keras. `https://github.com/fchollet/keras`.

Clement, R. and Sharp, D. (2003). Ngram and Bayesian Classification of Documents for Topic and Authorship. *Literary and Linguistic Computing*, 18(4):423–447.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience.

Coyotl-Morales, R. M., Villaseñor Pineda, L., Montes-y Gómez, M., and Rosso, P. (2006). Authorship attribution using word sequences. In *Proceedings of the 11th Iberoamerican Conference on Progress in Pattern Recognition, Image Analysis and Applications*, CIARP'06, pages 844–853, Berlin, Heidelberg. Springer-Verlag.

de Vel, O., Anderson, A., Corney, M., and Mohay, G. (2001). Mining e-mail content for author identification forensics. *ACM SIGMOD Record*, 30(4):55.

Devlin, J., Zbib, R., Huang, Z., Lamar, T., Schwartz, R., and Makhoul, J. (2014). Fast and robust neural network joint models for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1370–1380, Baltimore, Maryland. Association for Computational Linguistics.

Escalante, H. J., Solorio, T., and Montes-y Gomez, M. (2011). Local histograms of character n-grams for authorship attribution. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 288–298, Portland, Oregon, USA. Association for Computational Linguistics.

Forsyth, R. and Holmes, D. (1996). Feature-Finding for Text Classification. *Literary and Linguistic Computing*, 11(4):163–174.

Foster, D. (1989). *'Elegy' by W.S.: A Study in Attribution.* Associated University Press, Cranbury.

Fréry, J., Largeron, C., and Juganaru-Mathieu, M. (2014). UJM at CLEF in Author Identification—Notebook for PAN at CLEF 2014. In Cappellato, L., Ferro, N., Halvey, M., and Kraaij, W., editors, *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, 15-18 September, Sheffield, UK.* CEUR-WS.org.

Gamon, M. (2004). Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 611–617, Geneva, Switzerland. Association for Computational Linguistics.

Gatys, L. A., Ecker, A. S., and Bethge, M. (2015). A neural algorithm of artistic style. *CoRR*, abs/1508.06576.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *JMLR W&CP: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2010)*, volume 9, pages 249–256.

Goldberg, Y. and Hirst, G. (2017). *Neural Network Methods in Natural Language Processing.* Morgan & Claypool Publishers.

Gollub, T., Potthast, M., Beyer, A., Busse, M., Rangel, F., Rosso, P., Stamatatos, E., and Stein, B. (2013). Recent Trends in Digital Text Forensics and its Evaluation Plagiarism Detection , Author Identification , and Author Profiling. In *Proceeding of CLEF 2013*, pages 282–302.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning.* MIT Press. `http://www.deeplearningbook.org`.

Grant, T. D. (2007). Quantifying Evidence for Forensic Authorship Analysis. *International Journal of Speech, Language and Law*, 14(1):1–25.

Grieve, J. (2007). Quantitative Authorship Attribution: An Evaluation of Techniques. *Literary and Linguistic Computing*, 22(3):251–270.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235.

Guthrie, D. (2008). *Unsupervised Detection of Anomalous Text.* PhD thesis, University of Sheffield.

Halteren, H. V., Baayen, R. H., Tweedie, F., and Haverkort, M. (2005). New Machine Learning Methods Demonstrate the Existence of a Human Stylome. *Journal of Quantitative Linguistic*, 12(1):65–77.

Hirst, G. and Feiguina, O. (2007). Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts. *Literary and Linguistic Computing*, 22(4):405–417.

Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(2):107–116.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Holmes, D. (1992). A Stylometric Analysis of Mormon Scripture and Related Text. *Journal of the Royal Statistical Society Series A*, 155(1):91–120.

Hoover, D. (2003). Multivariate Analysis and the Study of Style Variation. *Literary and Linguistic Computing*, 18(4):341–360.

Hoover, D. L. (2004). Testing Burrows's Delta. *Literary and Linguistic Computing*, 19(4):453–475.

Houvardas, J. and Stamatatos, E. (2006). N-gram feature selection for authorship identification. In Euzenat, J. and Domingue, J., editors, *Artificial Intelligence: Methodology, Systems, and Applications*, pages 77–86, Berlin, Heidelberg. Springer Berlin Heidelberg.

Hovy, D. (2015). Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational*

Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 752–762, Beijing, China. Association for Computational Linguistics.

Huang, H.-H. and Chen, H.-H. (2011). Pause and stop labeling for chinese sentence boundary detection. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 146–153, Hissar, Bulgaria. RANLP 2011 Organising Committee.

Hürlimann, M., Weck, B., van den Berg, E., Šuster, S., and Nissim, M. (2015). GLAD: Groningen Lightweight Authorship Detection—Notebook for PAN at CLEF 2015. In Cappellato, L., Ferro, N., Jones, G., and San Juan, E., editors, *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France.* CEUR-WS.org.

Iqbal, F., Binsalleeh, H., Fung, B. C., and Debbabi, M. (2010). Mining writeprints from anonymous e-mails for forensic investigation. *Digital Investigation*, 7(1-2):56–64.

Jaffe, J. M., Lee, Y.-E., Huang, L., and Oshagan, H. (1995). Gender, Pseudonyms, and CMC: Masking Identities and Baring Souls. In *Proceeding of 45th Annual Conference of the International Communication Association*, Albuquerque, New Mexico, USA.

Jankowska, M., Kešelj, V., , and Milios, E. (2013). Proximity based One-class Classification with Common N-Gram Dissimilarity for Authorship Verification Task—Notebook for PAN at CLEF 2013. In Forner, P., Navigli, R., and Tufis, D., editors, *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain.*

Johannsen, A., Hovy, D., and Søgaard, A. (2015). Cross-lingual syntactic variation over age and gender. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 103–112, Beijing, China. Association for Computational Linguistics.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Juola, P. (2004). Ad-hoc authorship attribution competition. In *Proceedings of the Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing*, pages 175–176, Sweden.

Juola, P. (2008). Authorship Attribution. *Foundations and Trends® in Information Retrieval*, 1(3):233–334.

Juola, P. and Baayen, R. (2005). A controlled-corpus experiment in authorship identification by cross-entropy. *Literary and Linguistic Computing*, pages 1–10.

Juola, P. and Vescovi, D. (2010). Empirical evaluation of authorship obfuscation using JGAAP. In *Proceedings of the 3rd ACM workshop on Artificial intelligence and security - AISec '10*, page 14, New York, USA. ACM Press.

Jurafsky, D. and Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1st edition.

Kacmarcik, G. and Gamon, M. (2006). Obfuscating document stylometry to preserve author anonymity. In *Proceedings of the COLING/ACL*, pages 444–451, Morristown, NJ, USA. Association for Computational Linguistics.

Keselj, V., Peng, F., Cercone, N., and Thomas, C. (2003). N-gram-based author profiles for authorship attribution. In *Proceedings of the Pacific Association for Computational Linguistics*, pages 255–264, Halifax, Canada.

Kestemont, M., Luyckx, K., Daelemans, W., and Crombez, T. (2012). Cross-Genre Authorship Verification Using Unmasking. *English Studies*, 93(3):37–41.

Khmelev, D. and Teahan, W. (2003). A repetition based measure for verification of text collections and for text categorization. In *Proceedings of the 26th ACM SIGIR*, pages 104–110, Toronto, Canada.

Khonji, M. and Iraqi, Y. (2014). A Slightly-modified GI-based Author-verifier with Lots of Features (ASGALF)—Notebook for PAN at CLEF 2014. In Cappellato, L., Ferro, N., Halvey, M., and Kraaij, W., editors, *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, 15-18 September, Sheffield, UK*. CEUR-WS.org.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Kim, Y., Jernite, Y., Sontag, D., and Rush, A. M. (2016). Character-aware neural language models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 2741–2749. AAAI Press.

Kimler, M. (2003). *Using Style Markers for Detecting Plagiarism in Natural Language Documents*. PhD thesis, University of Skovde, Sweden.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Kjell, B. (1994). Discrimination of authorship using visualization. *Information Processing and Management*, 30(1):141–150.

Klarreich, E. (2003). Bookish math: Statistical tests are unraveling knotty literary mysteries. *Science News*, 164(25-26):392–392.

Klerke, S., Goldberg, Y., and Søgaard, A. (2016). Improving sentence compression by learning to predict gaze. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1528–1533. Association for Computational Linguistics.

Koppel, M., Akiva, N., and Dagan, I. (2006). Feature instability as a criterion for selecting potential style markers. *Journal of the American Society for Information Science and Technology*, 57(11):1519–1525.

Koppel, M., Argamon, S., and Gan, R. (2002). Automatically Categorizing Written Texts by Author Gender. *Journal of Literary and Linguistic Computing*, 17:401–412.

Koppel, M. and Schler, J. (2003). Exploiting stylistic idiosyncrasies for authorship attribution. In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, pages 69–72.

Koppel, M. and Schler, J. (2004). Authorship verification as a one-class classification problem. In *Machine Learning, Proceedings of the Twenty-first International Conference (ICML)*, Banff, Alberta, Canada.

Koppel, M., Schler, J., and Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1):9–26.

Koppel, M., Schler, J., and Argamon, S. (2011). Authorship attribution in the wild. *Language Resources and Evaluation*, 45(1):83–94.

Koppel, M., Schler, J., and Bonchek-Dokow, E. (2007). Measuring differentiability: Unmasking pseudonymous authors. *Journal of Machine Learning Research*, 8:1261–1276.

Koppel, M., Schler, J., and Zigdon, K. (2005). Automatically determining an anonymous author's native language. *Intelligence and Security Informatics*, pages 209–217.

Koppel, M. and Winter, Y. (2014). Determining if two documents are written by the same author. *Journal of the Association for Information Science and Technology*, 65(1):178–187.

Kukushkina, O. V., Polikarpov, A. A., and Khmelev, D. V. (2001). Using Literal and Grammatical Statistics for Authorship Attribution. 37(2):172–184.

Lambers, M. and Veenman, C. J. (2009). Forensic authorship attribution using compression distances to prototypes. In *Proceedings of The Third International Workshop on Computational Forensic (IWCF)*, pages 13–24, The Hague, The Netherlands. Springer Berlin Heidelberg.

Lee, W. S. and Liu, B. (2003). Learning with positive and unlabeled examples using weighted logistic regression. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, pages 448–455. AAAI Press.

Li, S., Chua, T.-S., Zhu, J., and Miao, C. (2016). Generative topic embedding: a continuous representation of documents. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 666–675, Berlin, Germany. Association for Computational Linguistics.

López-Monroy, A., y Gómez, M. M., Escalante, H., and Villaseñor-Pineda, L. (2014). Using Intra-Profile Information for Author Profiling—Notebook for PAN at CLEF 2014. In Cappellato, L., Ferro, N., Halvey, M., and Kraaij, W., editors, *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, 15-18 September, Sheffield, UK*. CEUR-WS.org.

Luong, M.-T., Le, Q. V., Sutskever, I., Vinyals, O., and Kaiser, L. (2016). Multitask sequence to sequence learning. In *Proceedings of International Conference on Learning Representations*, San Juan, Puerto Rico.

Luong, T., Kayser, M., and Manning, C. D. (2015). Deep neural language models for machine translation. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 305–309, Beijing, China. Association for Computational Linguistics.

Luong, T., Socher, R., and Manning, C. D. (2013). Better word representations with recursive neural networks for morphology. In *CoNLL*, pages 104–113, Sofia, Bulgaria.

Luyckx, K. and Daelemans, W. (2008). Authorship attribution and verification with many authors and limited data. In *Proceedings of the 22nd International Conference on Computational Linguistics - COLING '08*, volume 1, pages 513–520, Morristown, NJ, USA. Association for Computational Linguistics.

Luyckx, K. and Daelemans, W. (2010). The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing*, 26(1):35–55.

Martínez Alonso, H. and Plank, B. (2017). When is multitask learning effective? semantic sequence prediction under varying data conditions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 44–53, Valencia, Spain. Association for Computational Linguistics.

Marton, Y., Wu, N., and Hellerstein, L. (2005). On compression-based text classification. In *Proceedings of the European Conference on Information Retrieval*, pages 300–314, Santiago de Compostela, Spain.

Mayor, C., Gutierrez, J., Toledo, A., Martinez, R., Ledesma, P., Fuentes, G., , and Meza, I. (2014). A Single Author Style Representation for the Author Verification Task—Notebook for PAN at CLEF 2014. In Cappellato, L., Ferro, N., Halvey, M., and Kraaij, W., editors, *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, 15-18 September, Sheffield, UK*. CEUR-WS.org.

Mendenhall, T. (1887). The Characteristic Curves of Composition. *Science*, IX:37–49.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., and Khudanpur, S. (2010). Recurrent neural network based language model. *Interspeech*, 2:3.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013c). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Mikolov, T. and Zweig, G. (2012). Context dependent recurrent neural network language model. In *SLT*, pages 234–239. IEEE.

Mikros, G. K. and Argiri, E. K. (2007). Investigating topic influence in authorship attribution. In *Proceedings of the International Workshop on Plagiarism Analysis, Authorship Identification, and Near- Duplicate Detection*, pages 29–35.

Mikros, G. K. and Perifanos, K. A. (2013). Authorship Attribution in Greek Tweets Using Author ' s Multilevel N-Gram Profiles. In *2013 AAAI Spring Symposium*, pages 17–23, Washington, USA.

Mirowski, P. and Vlachos, A. (2015). Dependency recurrent neural language models for sentence completion. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 511–517, Beijing, China. Association for Computational Linguistics.

Modaresi, P. and Gross, P. (2014). A Language Independent Author Verifier Using Fuzzy C-Means Clustering—Notebook for PAN at CLEF 2014. In Cappellato, L., Ferro, N., Halvey, M., and Kraaij, W., editors, *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, 15-18 September, Sheffield, UK*. CEUR-WS.org.

Morton, A. Q. and Michaelson, S. (1990). The Qsum Plot. Technical report, University of Edinburgh.

Mosteller, F. and Wallace, D. L. (1964). Inference and Disputed Authorship: The Federalist. Addison Wesley.

Mulac, A., Studley, L., and Blau, S. (1990). The gender-linked language effect in primary and secondary students' impromptu essays. *Sex Roles*, 23.

Ney, H., Essen, U., and Kneser, R. (1994). On structuring probabilistic dependencies in stochastic language modelling. *Computer Speech and Language*, 8:1–38.

Oakes, M. (2004). Ant colony optimisation for stylometry: The federalist papers. In *Proceedings of the 5th International Conference on Recent Advances in Soft Computing*, pages 86–91.

Oliveira, W., Justino, E., and Oliveira, L. S. (2013). Comparing compression models for authorship attribution. *Forensic science international*, 228(1-3):100–4.

Panicheva, P., Cardiff, J., and Rosso, P. (2010). Personal Sense and Idiolect : Combining Authorship Attribution and Opinion Analysis. In *Proceeding of Seventh International Conference on Language Resources and Evaluation*, Malta.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peñas, A. and Rodrigo, A. (2011). A simple measure to assess non-response. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1415–1424.

Peng, F., Schuurmans, D., and Wang, S. (2004). Augmenting naive bayes classifiers with statistical language models. *Information Retrieval*, 7(3-4):317–345.

Peng, F., Schuurmanst, D., Kesel, V., and Wan, S. (2003). Language Independent Authorship Attribution using Character Level Language Models. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*, Budapest, Hungary.

Pennebaker, J. W., Booth, R. J., and Francis, M. E. (2001). *Linguistic Inquiry and Word Count: LIWC.* Erlbaum Publishers.

Plakias, S. and Stamatatos, E. (2008). Tensor space models for authorship identification. In *Proceedings of the 5th Hellenic Conference on Artificial Intelligence: Theories, Models and Applications*, SETN '08, pages 239–249, Berlin, Heidelberg. Springer-Verlag.

Potthast, M., Braun, S., Buz, T., Duffhauss, F., Friedrich, F., Gülzow, J., Köhler, J., Lötzsch, W., Müller, F., Müller, M., Paßmann, R., Reinke, B., Rettenmeier, L., Rometsch, T., Sommer, T., Träger, M., Wilhelm, S., Stein, B., Stamatatos, E., and Hagen, M. (2016a). Who Wrote the Web? Revisiting Influential Author Identification Research Applicable to Information Retrieval. In Ferro, N., Crestani, F., Moens, M.-F., Mothe, J., Silvestri, F., Di Nunzio, G., Hauff, C., and Silvello, G., editors, *Advances in Information Retrieval. 38th European Conference on IR Research (ECIR 16)*, volume 9626 of *Lecture Notes in Computer Science*, pages 393–407, Berlin Heidelberg New York. Springer.

Potthast, M., Hagen, M., and Stein, B. (2016b). Author Obfuscation: Attacking the State of the Art in Authorship Verification. In *Working Notes Papers of the CLEF 2016 Evaluation Labs*, volume 1609 of *CEUR Workshop Proceedings*. CLEF and CEUR-WS.org.

Rahimi, A., Baldwin, T., and Cohn, T. (2017). Continuous representation of location for geolocation and lexical dialectology using mixture density networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 167–176, Copenhagen, Denmark. Association for Computational Linguistics.

Rangel, F., Rosso, P., and Chugur, I. (2014). Overview of the 2nd Author Profiling Task at PAN 2014. In *Proceeding of CLEF 2014*.

Rao, J. R. and Rohatgi, P. (2000). Can Pseudonymity Really Guarantee Privacy. In *Proceedings of the 9th USENIX Security Symposium*, pages 85–96.

Rei, M. (2017). Semi-supervised multitask learning for sequence labeling. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 2121–2130.

Ribeiro, M., Singh, S., and Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

Rose, T., Stevenson, M., and Whitehead, M. (2002). The Reuters Corpus - from Yesterday's News to Tomorrow's Language Resources. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation, LREC 2002*, pages 827–832, Las Palmas, Canary Islands.

Rosso, P., Rangel, F., Potthast, M., Stamatatos, E., Tschuggnall, M., and Stein, B. (2016). Overview of PAN'16—New Challenges for Authorship Analysis: Cross-genre Profiling, Clustering, Diarization, and Obfuscation. In Fuhr, N., Quaresma, P., Larsen, B., Gonçalves, T., Balog, K., Macdonald, C., Cappellato, L., and Ferro, N., editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 7th International Conference of the CLEF Initiative (CLEF 16)*, Berlin Heidelberg New York. Springer.

Rudman, J. (1997). The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, pages 351–365.

Rudman, J. (2012). The State of Non-Traditional Authorship Attribution Studies—2012: Some Problems and Solutions. *English Studies*, 93(3):259–274.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523.

Sanderson, C. and Guenter, S. (2006). Short Text Authorship Attribution via Sequence Kernels , Markov Chains and Author Unmasking : An Investigation. In *EMNL*, number July, pages 482–491.

Sapkota, U., Bethard, S., Montes, M., and Solorio, T. (2015). Not all character n-grams are created equal: A study in authorship attribution. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–102, Denver, Colorado. Association for Computational Linguistics.

Sapkota, U., Solorio, T., Montes, M., Bethard, S., and Rosso, P. (2014). Cross-topic authorship attribution: Will out-of-topic data help? In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1228–1237, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Sari, Y., Vlachos, A., and Stevenson, M. (2017). Continuous n-gram representations for authorship attribution. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 267–273, Valencia, Spain. Association for Computational Linguistics.

Savoy, J. (2013). Feature selections for authorship attribution. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, pages 929–941.

Schler, J., Koppel, M., Argamon, S., and Pennebaker, J. W. (2006). Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 199–205. AAAI.

Schwartz, R., Tsur, O., Rappoport, A., and Koppel, M. (2013). Authorship Attribution of Micro-Messages. In *2013 Conference on Empirical Methods in Natural Language Processing*, number October, pages 1880–1891, Seattle, USA.

Schwenk, H., Dchelotte, D., and Gauvain, J.-L. (2006). Continuous space language models for statistical machine translation. In *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, COLING-ACL '06, pages 723–730, Stroudsburg, PA, USA. Association for Computational Linguistics.

Seidman, S. (2013). Authorship Verification Using the Impostors Method—Notebook for PAN at CLEF 2013. In Forner, P., Navigli, R., and Tufis, D., editors, *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain.*

Seroussi, Y., Smyth, R., and Zukerman, I. (2011). Ghosts from the high court's past: Evidence from computational linguistics for dixon ghosting for mctiernan and rich. *University of New South Wales Law Journal*, 34(3):984–1005.

Seroussi, Y., Zukerman, I., and Bohnert, F. (2010). Collaborative inference of sentiments from texts. In *Proceedings of 18th International Conference on User Modeling, Adaptation, and Personalization (UMAP)*, pages 195–206, Big Island, HI, USA. Springer Berlin Heidelberg.

Seroussi, Y., Zukerman, I., and Bohnert, F. (2013). Authorship Attribution with Topic Models. *Journal Computational Linguistics*, 40(2):269–310.

Shrestha, P., Sierra, S., Gonzalez, F., Montes, M., Rosso, P., and Solorio, T. (2017). Convolutional neural networks for authorship attribution of short texts.

In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 669–674, Valencia, Spain. Association for Computational Linguistics.

Smith, M. W. A. (1983). Recent Experience and New Developments of Methods for the Determination of Authorship. *Association for Literary and Linguistic Computing Bulleting*, 11:73–82.

Søgaard, A. and Goldberg, Y. (2016). Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 231–235, Berlin, Germany. Association for Computational Linguistics.

Stamatatos, E. (2006). Ensemble-based author identification using character n-grams. In *Proceedings of the 3rd International Workshop on Text-based Information Retrieval*, pages 41–46.

Stamatatos, E. (2008). Author identification: Using text sampling to handle the class imbalance problem. *Information Processing and Management*, 44(2):790 – 799.

Stamatatos, E. (2009a). A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

Stamatatos, E. (2009b). Intrinsic Plagiarism Detection Using Character n-gram Profiles. In Stein, B., Rosso, P., Stamatatos, E., Koppel, M., and Agirre, E., editors, *SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09)*, pages 38–46. Universidad Politécnica de Valencia and CEUR-WS.org.

Stamatatos, E. (2013). On the Robustness of Authorship Attribution Based on Character n-gram Features. *Journal of Law and Policy*, 21(2):421–439.

Stamatatos, E., amd Ben Verhoeven, W. D., Juola, P., López-López, A., Potthast, M., and Stein, B. (2015). Overview of the Author Identification Task at PAN 2015. In Cappellato, L., Ferro, N., Jones, G., and San Juan, E., editors, *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France.* CEUR-WS.org.

Stamatatos, E., Daelemans, W., Verhoeven, B., Potthast, M., Stein, B., Juola, P., Sanchez-Perez, M., and Barrón-Cedeño, A. (2014). Overview of the Author

Identification Task at PAN 2014. In Cappellato, L., Ferro, N., Halvey, M., and Kraaij, W., editors, *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, 15-18 September, Sheffield, UK*. CEUR-WS.org.

Stamatatos, E., Fakotakis, N., and Kokkinakis, G. (2001). Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, pages 193–214.

Stamatatos, E., Kokkinakis, G., and Fakotakis, N. (2000). Automatic text categorization in terms of genre and author. *Comput. Linguist.*, 26(4):471–495.

Stein, B. and Meyer Zu Eissen, S. (2007). Intrinsic plagiarism analysis with meta learning. In *CEUR Workshop Proceedings*, volume 276, pages 45–50.

Stolcke, A. (2002). Srilm – an extensible language modeling toolkit. In *Proceedings of The 7th International Conference on Spoken Language Processing (ICSLP)*, pages 901–904, Denver, Colorado, USA.

Sundermeyer, M., Schluter, R., and Ney, H. (2012). Lstm neural networks for language modeling. In *Interspeech*, pages 194–197, Portland, OR, USA.

Sutskever, I. (2013). *Training Recurrent Neural Networks*. PhD thesis, Toronto, Ont., Canada, Canada. AAINS22066.

Swinson, T. and Reyna, C. (2013). Authorship Attribution Using Stopword Graphs. pages 1–9.

Tallec, C. and Ollivier, Y. (2017). Unbiasing truncated backpropagation through time. *CoRR*, abs/1705.08209.

Tschuggnall, M., Stamatatos, E., Verhoeven, B., Daelemans, W., Specht, G., Stein, B., and Potthast, M. (2017). Overview of the Author Identification Task at PAN-2017: Style Breach Detection and Author Clustering. In Cappellato, L., Ferro, N., Goeuriot, L., and Mandl, T., editors, *Working Notes Papers of the CLEF 2017 Evaluation Labs*, volume 1866 of *CEUR Workshop Proceedings*. CLEF and CEUR-WS.org.

van Halteren, H. (2004). Linguistic profiling for authorship recognition and verification. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 199–206, Barcelona, Spain.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2014). Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555.

Wang, S. and Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, pages 90–94, Stroudsburg, PA, USA. Association for Computational Linguistics.

Weinberger, K., Dasgupta, A., Langford, J., Smola, A., and Attenberg, J. (2009). Feature hashing for large scale multitask learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1113–1120, New York, NY, USA. ACM.

Werbos, P. J. (1990). Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560.

Williams, C. B. (1940). A note on the statistical analysis of sentence-length as a criterion of literary style. *Biometrika*, 31(3):63–90.

Williams, R. J. and Peng, J. (1990). An efficient gradient-based algorithm for on-line training of recurrent network trajectories. *Neural Computation*, 2(4):490–501.

Yannakoudakis, H., Rei, M., Andersen, Ø. E., and Yuan, Z. (2017). Neural sequence-labelling models for grammatical error correction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2795–2806, Copenhagen, Denmark. Association for Computational Linguistics.

Yule, G. (1939). On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship. *Biometrika*, 30(3):363–390.

Yule, G. (1944). The statistical study of literary vocabulary. *The Modern Language Review*, 39(3):291–293.

Zaremba, W., Sutskever, I., and Vinyals, O. (2014). Recurrent neural network regularization. *CoRR*, abs/1409.2329.

Zechner, M., Muhr, M., Kern, R., and Graz, K.-c. (2009). External and Intrinsic Plagiarism Detection. In *Proc. of 3rd Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse and 1st International Competition on Plagiarism Detection*, pages 47—-55.

Zhang, X., Zhao, J., and LeCun, Y. (2015). Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 649–657, Montreal, Canada. MIT Press.

Zhao, Y. and Zobel, J. (2005). Effective and scalable authorship attribution using function words. *Information Retrieval Technology*, pages 174–189.

Zhao, Y. and Zobel, J. (2007). Entropy-based authorship search in large document collections. In *Proceedings of the 29th European Conference on IR Research*, ECIR'07, pages 381–392, Rome, Italy. Springer-Verlag.

Zheng, R., Li, J., Chen, H., and 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.

Zimmer, B. (2013). The science that uncovered J.K. Rowling's literary hocus-pocus. *The Wall Street Journal*.

Zipf, G. K. (1933). Selected Studies of the Principle of Relative Frequency in Language. *Language*, 9(1):89–92.

Zweig, G., Platt, J. C., Meek, C., Burges, C. J., Yessenalina, A., and Liu, Q. (2012). Computational approaches to sentence completion. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 601–610, Jeju Island, Korea. Association for Computational Linguistics.

Appendix A

A.1 Predefined Function Words and Punctuation

a about above after again against all am an and any are aren’t as at be because been before being below between both but by can’t cannot could couldn’t did didn’t do does doesn’t doing don’t down during each few for from further had hadn’t has hasn’t have haven’t having he he’d he’ll he’s her here here’s hers herself him himself his how how’s i i’d i’ll i’m i’ve if in into is isn’t it it’s its itself let’s me more most mustn’t my myself no nor not of off on once only or other ought our ours ourselves out over own same shan’t she she’d she’ll she’s should shouldn’t so some such than that that’s the their theirs them themselves then there there’s these they they’d they’ll they’re they’ve this those through to too under until up very was wasn’t we we’d we’ll we’re we’ve were weren’t what what’s when when’s where where’s which while who who’s whom why why’s with won’t would wouldn’t you you’d you’ll you’re you’ve your yours yourself yourselves

Figure A.1: Predefined function words

“’, ‘:’, ‘,’, ‘_’, ‘!’, ‘?’, ‘;’, ‘.’, “‘’, ‘(’, ‘)’, ‘-’

Figure A.2: Predefined punctuation

A.2 The Blog Authorship Corpus Details

These following tables present the statistic details of The Blog Authorship Corpus according to age and gender profiles. The tables are presented without modification from **?** work.

feature	10s	20s	30s
maths	**1.05±0.06**	0.03±0.00	0.02±0.01
homework	**1.37±0.06**	0.18±0.01	0.15±0.02
bored	**3.84±0.27**	1.11±0.14	0.47±0.04
sis	**0.74±0.04**	0.26±0.03	0.10±0.02
boring	**3.69±0.10**	1.02±0.04	0.63±0.05
awesome	**2.92±0.08**	1.28±0.04	0.57±0.04
mum	**1.25±0.06**	0.41±0.04	0.23±0.04
crappy	**0.46±0.02**	0.28±0.02	0.11±0.01
mad	**2.16±0.07**	0.80±0.03	0.53±0.04
dumb	**0.89±0.04**	0.45±0.03	0.22±0.03
semester	0.22±0.02	**0.44±0.03**	0.18±0.04
apartment	0.18±0.02	**1.23±0.05**	0.55±0.05
drunk	0.77±0.04	**0.88±0.03**	0.41±0.05
beer	0.32±0.02	**1.15±0.05**	0.70±0.05
student	0.65±0.04	**0.98±0.05**	0.61±0.06
album	0.64±0.05	**0.84±0.06**	0.56±0.08
college	1.51±0.07	1.92±0.07	1.31±0.09
someday	0.35±0.02	**0.40±0.02**	0.28±0.03
dating	0.31±0.02	**0.52±0.03**	0.37±0.04
bar	0.45±0.03	**1.53±0.06**	1.11±0.08
marriage	0.27±0.03	0.83±0.05	**1.41±0.13**
development	0.16±0.02	0.50±0.03	**0.82±0.10**
campaign	0.14±0.02	0.38±0.03	**0.70±0.07**
tax	0.14±0.02	0.38±0.03	**0.72±0.11**
local	0.38±0.02	1.18±0.04	**1.85±0.10**
democratic	0.13±0.02	0.29±0.02	**0.59±0.05**
son	0.51±0.03	0.92±0.05	**2.37±0.16**
systems	0.12±0.01	0.36±0.03	**0.55±0.06**
provide	0.15±0.01	0.54±0.03	**0.69±0.05**
workers	0.10±0.01	0.35±0.02	**0.46±0.04**

Table A.1: Word frequency (per 10,000 words) and standard error by age

feature	male	female
linux	**0.53±0.04**	0.03±0.01
microsoft	**0.63±0.05**	0.08±0.01
gaming	**0.25±0.02**	0.04±0.00
server	**0.76±0.05**	0.13±0.01
software	**0.99±0.05**	0.17±0.02
gb	**0.27±0.02**	0.05±0.01
programming	**0.36±0.02**	0.08±0.01
google	**0.90±0.04**	0.19±0.02
data	**0.62±0.03**	0.14±0.01
graphics	**0.27±0.02**	0.06±0.01
india	**0.62±0.04**	0.15±0.01
nations	**0.25±0.01**	0.06±0.01
democracy	**0.23±0.01**	0.06±0.01
users	**0.45±0.02**	0.11±0.01
economic	**0.26±0.01**	0.07±0.01
shopping	0.66±0.02	**1.48±0.03**
mom	2.07±0.05	**4.69±0.08**
cried	0.31±0.01	**0.72±0.02**
freaked	0.08±0.01	**0.21±0.01**
pink	0.33±0.02	**0.85±0.03**
cute	0.83±0.03	**2.32±0.04**
gosh	0.17±0.01	**0.47±0.02**
kisses	0.08±0.01	**0.28±0.01**
yummy	0.10±0.01	**0.36±0.01**
mommy	0.08±0.01	**0.31±0.02**
boyfriend	0.41±0.02	**1.73±0.04**
skirt	0.06±0.01	**0.26±0.01**
adorable	0.05±0.00	**0.23±0.01**
husband	0.28±0.01	**1.38±0.04**
hubby	0.01±0.00	**0.30±0.02**

Table A.2: Word frequency (per 10,000 words) and standard error by gender

		10s	20s	30s	all
pronouns	all	**1316.7**	1173.7	1104.4	
	male	1216.4	1063.0	968.7	1113.8
	female	1416.9	1284.5	1240.1	**1334.1**
assent	all	**33.7**	20.1	17.0	
	male	30.0	18.5	15.3	22.9
	female	37.5	21.7	18.7	**28.0**
negation	all	**162.0**	157.5	149.3	
	male	153.4	146.7	137.8	148.1
	female	170.7	168.4	160.8	**168.2**
determiners	all	488.9	619.9	671.5	
	male	542.1	661.9	715.4	**619.1**
	female	435.7	578.0	627.6	525.0
prepositions	all	1077.0	1231.9	**1276.6**	
	male	1123.5	1250.8	1296.7	**1203.6**
	female	1030.5	1212.9	1256.5	1141.8
blogwords	all	**122.1**	34.8	20.4	
	male	99.2	31.3	18.7	58.3
	female	145.1	38.4	22.1	**81.4**
hyperlinks	all	**20.7**	35.0	38.8	
	male	25.4	41.7	49.1	35.9
	female	16.0	28.4	28.6	**23.1**
post length	all	195.0	210.0	**221.0**	
	male	191.4	2017.5	204.1	201.0
	female	198.8	213.6	240.3	**213.0**

Table A.3: Frequency (per 10,000 words) of stylistic features per gender per age bracket