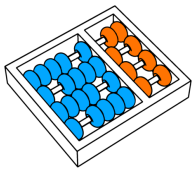
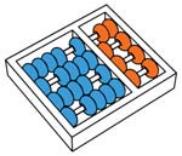


Thiago Cavalcante

"Authorship Attribution on Micro-Messages."

"*Atribuição de Autoria em Micro-Mensagens.*"

CAMPINAS
2014



University of Campinas
Institute of Computing

Universidade Estadual de Campinas
Instituto de Computação

Thiago Cavalcante

"Authorship Attribution on Micro-Messages."

Supervisor: *Orientador(a):* **Prof. Dr. Ariadne Maria Brito Rizzoni Carvalho**

Co-Supervisor: *Co-orientador(a):* **Prof. Dr. Anderson de Rezende Rocha**

"Atribuição de Autoria em Micro-Mensagens."

MSc Dissertation presented to the Post Graduate Program of the Institute of Computing of the University of Campinas to obtain a Master degree in Computer Science.

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Computação da Universidade Estadual de Campinas para obtenção do título de Mestre em Ciência da Computação.

This volume corresponds to the final version of the Dissertation defended by Thiago Cavalcante, under the supervision of Prof. Dr. Ariadne Maria Brito Rizzoni Carvalho.

Este exemplar corresponde à versão final da Dissertação defendida por Thiago Cavalcante, sob orientação de Prof. Dr. Ariadne Maria Brito Rizzoni Carvalho.

Ariadne M. B. R. Carvalho

Supervisor's signature / *Assinatura do Orientador(a)*

CAMPINAS
2014

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Maria Fabiana Bezerra Muller - CRB 8/6162

C314a

Cavalcante, Thiago, 1989-
Authorship attribution on micro-messages / Thiago Cavalcante. – Campinas, SP : [s.n.], 2014.

Orientador: Ariadne Maria Brito Rizzoni Carvalho.
Coorientador: Anderson de Rezende Rocha.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Computação.

1. Autoria. 2. Redes sociais. 3. Blogs. 4. Big data. 5. Internet. 6. Comunicação na tecnologia. I. Carvalho, Ariadne Maria Brito Rizzoni,1958-. II. Rocha, Anderson de Rezende,1980-. III. Universidade Estadual de Campinas. Instituto de Computação. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: Atribuição de autoria em micro-mensagens
Palavras-chave em inglês:
Authorship
Social networks
Blogs
Big data
Internet
Communication in technology
Área de concentração: Ciência da Computação
Titulação: Mestre em Ciência da Computação
Banca examinadora:
Ariadne Maria Brito Rizzoni Carvalho [Orientador]
Hélio Pedrini
Cinthia Obladen de Almendra Freitas
Data de defesa: 15-12-2014
Programa de Pós-Graduação: Ciência da Computação

TERMO DE APROVAÇÃO

Defesa de Dissertação de Mestrado em Ciência da Computação, apresentada pelo(a) Mestrando(a) **Thiago Cavalcante**, aprovado(a) em **15 de dezembro de 2014**, pela Banca examinadora composta pelos Professores(as) Doutores(as):

Prof(a). Dr(a). Cinthia Obladen de Almendra Freitas
Titular

Prof(a). Dr(a). Hélio Pedrini
Titular

Prof(a). Dr(a). Ariadne Maria Brito Rizzoni Carvalho
Presidente

Institute of Computing /*Instituto de Computação*
University of Campinas /*Universidade Estadual de Campinas*

Authorship Attribution on Micro-Messages.

Thiago Cavalcante[1]

December 15, 2014

Examiner Board/*Banca Examinadora*:

- Prof. Dr. Ariadne Maria Brito Rizzoni Carvalho (Supervisor/*Orientador*)
- Prof. Dr. Hélio Pedrini
  Institute of Computing - UNICAMP
- Prof. Dr. Cinthia Obladen de Almendra Freitas
  Pontifícia Universidade Católica do Paraná - PUC/PR
- Prof Dr. Neucimar Jerônimo Leite
  Institute of Computing - UNICAMP (Substitute/*Suplente*)
- Dr. Norton Trevisan Roman
  Escola de Artes, Ciências e Humanidades - EACH/USP (Substitute/*Suplente*)

[1]Financial support: CAPES scholarship (process #01P45543013) 2012–2014

Abstract

With the ever-growing use of social media, authorship attribution plays an important role in avoiding cybercrime, and helping the analysis of online trails left behind by cyber pranks, stalkers, bullies, identity thieves and alike. In this dissertation, we propose a method for authorship attribution in micro blogs with efficiency one hundred to a thousand times faster than state-of-the-art counterparts. The method relies on a powerful and scalable feature representation approach taking advantage of user patterns on micro-blog messages, and also on a custom-tailored pattern classifier adapted to deal with big data and high-dimensional data. Finally, we discuss search-space reduction when analysing hundreds of online suspects and millions of online micro messages, which makes this approach invaluable for digital forensics and law enforcement.

Resumo

Com o crescimento continuo do uso de midias sociais, a atribuição de autoria tem um papel imortante na prevenção dos crimes cibernéticos e na análise de rastros online deixados por assediadores, *bullies*, ladrões de identidade entre outros. Nesta dissertação, nós propusemos um método para atribuição de autoria que é de cem a mil vezes mais rápido que o estado da arte. Nós também obtivemos uma acurácia 65% na classificação de 50 autores. O método proposto se baseia numa representação de caracteristicas escalável utilizando os padrões das mensagens dos micro-blogs, e também nos utilizamos de um classificador de padrões customizado para lidar com grandes quantidades de dados e alta dimensionalidade. Por fim, nós discutimos a redução do espaço de busca na análise de centenas de suspeitos online e milões de micro mensagens online, o que torna essa abordagem valiosa para forense digital e aplicação das leis.

In memory of my father.

Acknowledgements

First of all, I would like to thank God, who created a universe in such a way that we are able to find fixed patterns, some of these which are objects of my studies. I also thank my mother, father, sister and my fianceé who always supported, encouraged and helped me to pursue my dreams.

My gratitude also goes to my advisors Ariadne and Anderson, who were with me through this important step in my life, not only with scientific but also with personal advices that made the going ever easier. Also, to all my teachers whom guided me through knowledge, moral and wisdom from the begining of my life until now.

Last, but not least, to all of my family and friends who were with me for the jorney sometimes with helpful advices and sometimes providing entertainment and laughter. And also, I am grateful to all those who made this work possible from the founders of our educational institutions, through our governors and administrators to the janitors and security personnel – which not only do their jobs, but are also a pleasant company for an evening conversation.

Thank you all, for making this possible.

"A man does not call a line crooked unless he has some idea of a straight line. What was I comparing this universe with when I called it unjust?"

C. S. Lewis

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The recent explosion of social media brings about a great deal of freedom of speech, but this often comes along with anonymity. The problem is even more complex with deep web, the World Wide Web content that is not directly indexed by search engines. With so much freedom and possibilities of social connections, it is inevitable that people start using anonymity as a weapon for their personal agenda, such as defaming politicians with opposite views, or impersonating one another. Recent media articles abound discussing online harassment cases, identity theft, online impersonation, stalking and alike to name just a few [98, 104]. A recent survey from the Wall Street Journal in the United States revealed that in 2011, more than 5% of Facebook, 6.3% of Twitter, and 7% of Google+ users suffered from identity theft [121]. The most alarming reality is that these numbers were up 13% with respect to the previous year.

The possibility of anonymity raises the need to identify authors by other means. The user messages themselves can be used for the task. Notwithstanding when it comes to social media, we have to make sense of a massive amount of data, transform it into information and, ultimately, into knowledge. Improved learning solutions are needed for gleaning knowledge and insights from such data. To complicate matters even more, each message is, by itself, too small to allow the exploration of patterns and user trails. In this dissertation, we analyze micro-blog messages, more specifically Twitter data, whereby 500+ millions of new messages are exchanged everyday [64]. Easily, a cybercrime committed either by means of Twitter messages or committed elsewhere but with evidences on it, might have a large number of suspects, generating an even bigger amount of data to be analyzed and a real challenge for authorship

attribution.

1.1 Problem

Given this context, we ought to find means to identify the authors of such content. Surely there are many ways to find out the author, which include IP addresses, traditional investigation or even the user account itself. The problem lies whenever these methods fail; many services, widely available, can hide the user IP address such as onion routing[1] [99].

The goal of authorship attribution is to identify authors of texts through features from the style of the author alone; this is called stylometry. The longer the text is, the easier to capture such features and more reliable they become. When dealing with micro-blog messages, the texts are very short and therefore a smaller set of stylometric features are present in each one. Some authors even suggest joining the messages in a single document [86]. Even with preliminary results showing some improvement, this is not realistic since we may not have more than one message we wish to know the author [103], and whenever dealing with anonymous messages we cannot guarantee that all of them belong to the same author.

Differently from essays and books, the nature of web-based messages include internet meta-language such as differentiated punctuation, emoticons[2], internet expressions[3] and the lack of spell-checking. These differentiated characteristics seem to be key for attribution in the context of micro-messages. Notwithstanding they also raise another challenge, since there are thousands of different punctuation uses, emoticons, expressions and grammar errors and a gigantic number of features generated by them.

To adapt authorship attribution to this media, we need to find stylometric features that capture the web-based features. Such features are generally sparse in nature, generating huge feature vectors. Since messages are short, the task requires a great amount of training data to increase the classification robustness, as well as

[1] Onion Routing is a network where the user data will be encrypted and sent back and forth trough other peers on the network slowing down traffic, but making the user virtually anonymous.

[2] Emoticons are faces made with ascii symbols such as: ":-)", ":'(", ":-D" and "D-:"

[3] It is very common that an expression or abbreviation becomes very popular in messages and social media around the web; common examples are: lol = laugh out loud; brb = be right back; cya = see you.

faster and scalable classifiers to deal with the amount of data and features which are generated.

1.2 Forensic Application

Given the problem of authorship identification, it seems clear that these techniques have a direct appeal in investigations and trials. Although the most promising works in authorship attribution show results far from perfect, these techniques have been, and should continue to be used to give directions in investigations, possibly narrowing it down to a few suspects or backing up evidence against a suspect.

A famous example is Ted Kazinsky, the unabomber case in the United States. His brother recognized the writing style of the manifesto and tests pointed to him. Although this evidence was never used in court, it was the starting point for the investigations. With the fast growing of social media, the production of web texts increased drastically and with it the infamous cybercrime. More than ever, we need to deal with the increasing number [87] of anonymous cybercrimes and the volume of data to be analyzed in such situations.

In this work, we suggest the same kind of application. Authorship attribution is not precise as DNA testing, and will rarely be used in a courtroom by itself. But as we show further on, we can greatly reduce the amount of suspects in a case, therefore being extremely useful as a starting point in a forensic investigation. Moreover, we explored new classifiers that deal with the big data generated by social media.

1.3 Working Hypothesis

Our working hypothesis is that even with micro-blog texts being much smaller than books or essays, it is still possible to perform authorship attribution using a dynamic set of features, adaptable to each group of users.

When working with features adaptable to each set of users, we also have to deal with a large number of features which are generated; therefore, we also need to use these feature vectors in scalable classifiers to deal with the problem.

1.4 Contribution

First, we approach the problem proposing a characterization technique that captures character and word properties that are used (character n-grams and word n-grams). These features were explored by few sources in this context because often they lead to high-dimensional feature vectors unsuitable to most classification techniques. Such vectors represent a real challenge for traditional classifiers to carve decision surfaces without being doomed by the curse of dimensionality. Differently, here we address this problem by showing an alternative for the traditional Support Vector Machine (SVM) classifier, which better deals with big data in terms of accuracy and performance. We rely upon Power Mean SVM [123], a solution recently proposed for large-scale visual classification and herein instantiated with success for the problem of large-scale authorship attribution of micro-blog messages. Finally, we discuss the viability of reducing the search space for forensic analysis when dealing with hundreds of suspects and millions of messages.

1.5 Dissertation Organization

The remainder of this dissertation is organized as follows, Chapter 2 contains a survey of the state-of-the-art techniques in authorship attribution, from the early works in authorship to the more modern works in micro-blog authorship attribution. In Chapter 3, we present the methodology for authorship attribution proposed in this dissertation and, in Chapter 4, we show the experimental results and present a discussion about the meaning of each one. In Chapter 5, we draw conclusions about the work, and discuss a general guideline for future development in authorship atributtion in micro-blogs.

Chapter 2

State of the art

As soon as humans started writing, the need to know who wrote what appeared instantly. This need makes authorship attribution a very old field of study. Orders from kings, army generals and even from family elders date as old as the society itself. Curiously enough humans are very adaptive in perceiving the authors of a message. Take for example a small child approached by a stranger on behalf of its mother[1]:

> Stranger: "Hey, your mom asked me to take you home"
>
> Boy: "She did? But what did she say?"
>
> Stranger: "She said: Please take my son, Michael, home"
>
> Boy: "My mom always calls me Mike, go away before I call the police!"

Although this scene does not come from a real scenario, it is natural for us humans to be able to recognize patterns in the speeches of those close to us. This means that we do have patterns in our ways to comunicate, which have been an object of study since words themselves were created. Patrick Juola extracts an instance of authorship from the bible, dating approximatelly 1,000 b.C., where the Israelites could identify a spy by a word he could not spell properlly [50].

Even older texts such as the code of Hamurabi, which is considered the oldest text of significant length and which dates from aproximatelly 1,780 B.C. says:

[1]This text was created by the author as a mere illustration.

> "If a judge try a case, reach a decision, and present his judgement in writing; if later error shall appear in his decision, and it be through his own fault, then he shall pay twelve times the fine set by him in the case, and he shall be publicly removed from the judge's bench, and never again shall he sit there to render judgement."

The author of a given text, in this case a judge, is of ultmost importance since early times. Making authorship attribution perhaps one of the oldest fields of study with many examples throughout the history, such as the Donation of Constantine[2], which was proved to be a forgery by Lorenzzo de Valla in 1439. And specially with the renaiscense age and its many books, doubt was casted uppon the real authors of many works of this and older historical periods.

Authorship attribution was mostly done by linguistic experts who would identify traits of an author, and try to answer if he did or did not write a particular text. In our work, we are interested in other methods of identifying an author that do not rely on human experts. Statistical and computational methods over metrics in the text are called stylometry. It is used to capture the style of each author and classify texts accordingly.

2.1 Early Authorship Attribution

Although authorship attribution was always important, it was more of a detective work; in order to find an unknown essay author, we would look for clues, words an author never uses, mistakes he always makes, ideology, amongst other features. The field was brought to a more mathematical approach in 1887 when Mendenhall [79] experimented with books from three different authors in an attempt to identify an author by the sentence and word lengths.

Mendenhall findings were based on a letter of the English logician Augustus de Morgan [27], who suggested to a friend, early in 1851, that the authorship could be stablished by analysing by the length of someone's text [50]. Although this approach would be proven inefficient later on, it was the first attempt to find a set of metrics that would identify an author instead of looking for personal traits.

[2]The Donation of Constantine was a document in wich the emperor Constantine I gives control of a large portion of the west Roman empire to the Pope Sylvester I.

Right after, in 1888, Mascol [75] would take a different approach on the subject with the aid of function words and the most frequent punctuation in the Pauline Epistles. This was the first work to look for function words in authorship attribution. Although nowadays it is considered to be a very good stylometric characteristic, it would only be used again almost a century later, in 1962, in the work of Ellegaard [31].

Many years later, in 1938 and 1944 respectively, Yule would suggest using average sentence length in characters [124] and vocabulary richness (k-measure) [125] as new measures due to the inefficiency of the methods proposed thus far. Although still inefficient, the use of average sentence length tries to round up patterns found in the document as a whole instead of using isolated sentences as a metric. However, the vocabulary richness – also known as K-measure – was done by statistical analysis on the frequency of the words and highly based on Zipf law[3] [130], therefore, a more robust metric than simple lengths.

These were followed by many authors using the same metrics to discover authorship, and in different languages such as German [33] and Greek [120, 88]. But those methods were based on word and sentence length, which future studies would show that are highly unstable among different texts from the same author [50].

It is also worthy noting that these early methods were based on direct comparison of the data generated; there was no classifier or statistical technique involved in the comparisons whatsoever. Table 2.1 summarizes the early works in authorship attribution and the features used.

[3]Zipf law states that the frequency of a word in a corpus is inversely proportional to its rank in the frequency table by an exponential factor.

Table 2.1: Works in modern authorship attribution.

Source	**Features Used**	**Classifier**	**Corpus**
Mendenhall 1887 [79]	Word and Sentence Length	-	Bacon, Marlow and Shakespeare
Mascol 1888 [75]	Function Words and Punctuation	-	Pauline Epistles
Yule 1938 [124]	Sentence Length	-	de Gerson
Yule 1944 [125]	K-measure	-	de Gerson
Fucks 1952 [33]	Word Length	-	English and german authors
Wake 1957 [120]	Words per sentence	-	Greek authors
Ellegaard 1962 [31]	Function Words	-	The Junius letters
Morton 1965 [88]	Sentence Length	-	Greek authors

2.2 Modern Authorship Attribution

In 1964, Mosteller and Wallace [93] started a study over the Federalist Papers[4] bringing some statistical tools to the play. They analyzed 10 function words, as suggested in the work of Ellegaard [31], with a Naïve-Bayes classifier to predict the author of each one of the disputed papers.

It is consensus amongst authors that the work of Mosteller and Wallace is one of the most influential in this area [38, 50, 61, 112]. Not only it introduced the use of classifiers in the authorship attribution methods, but it also changed the stylometric approach for a more robust one than simple sentence and word lengths by using the frequencies of the function words [50].

That would begun a new age on the authorship attribution process. The main reason being that the results from their work conformed to the trending analysis from the expert linguistics. Also there were some notable results such as the word "upon" appearing on average 3.24 times for each 1,000 words in the works of Hamilton, and only 0.23 in the works of Madison [38].

Although it is important to remember that the experimental setup with the federalist papers is not good for the simple reason that the real author is unknown, so it is impossible to determine if the method used was correct, we can only know if it conforms to the opinion of the experts. The correct approach would be testing the method for texts whose authorship is well known first. This same setup would be repeated for many other authors.

Unfortunately, following this great success, stylometry would struggle for almost two decades due to a lack of cohesion in the field. That is specially evident in the works of Merriam [80, 81, 82], who used the methods proposed by Morton [89], receiving much criticism in the works of Smith [107, 106], who would also criticize another work by Morton [90, 108, 83].

These struggles seen by the eyes of humanist scholars would cast doubt on the use of statistical methods to attribute authorship. Stylometry would be stuck until the early 90's when John Burrows [15, 16, 18, 17] proposed the introduction of a multivariate analysis in stylometry. Basically, he generated a vector of frequencies of function words and applied Principal Component Analysis (PCA) over it and classified the authors by simple clusterization.

[4]Federalist Papers are a collection of essays promoting the ratification of the US Constitution, some of them claimed by Alexander Hamilton, James Madison and John Jay [93].

The multivariate analysis was so successful that it became well established in the authorship attribution. Many authors wrote only asserting the success of the method, such as Holmes and Forsyth [39], Baayen et al. [8] and Tweedie [117]. In the subsequent years, many authors used these methods to analyze the authorship of many texts like the Bible gospels [78, 77], Shakespeare [68], Goldsmith [30] and many other classical texts [116, 73].

By the same time, another analysis would generate an enormous controversy in authorship attribution, commonly called the Cumulative Sum (CUSUM) controversy [38, 50]. CUSUM charts were used primarily in quality control; in 1991, Morton published a method involving those [92, 91], and it would rapidly gain notoriety and even started being used in courtrooms.

Morton's technique was based on the assumption that a person has habits with words which repeat themselves. The main metrics in his works were the use of short words (words with less than three letters), vowel words (words beginning with vowels) and the joint of the two, short+vowel words [38].

Many authors from the academic community criticized the use of CUSUM charts as forensic evidence [22, 26, 35, 34, 36, 102, 42]. Also BBC television challenged Morton's technique live on air and it was unable to distinguish the writings of the chief of justice in England and a convicted felon [50].

In spite of the controversies there are still some works that use alternative variations of the CUSUM method such as the Weighted Cumulative Sum (WCUSUM) [13, 109, 110].

In the coming years, authorship attribution would be seen as a problem of pattern recognition. With the trending of neural networks in the early 90's, this classifier would become very popular in authorship attribution, making its debut in 1994 in a work by Matthews and Merriam [76] and followed by many others [56, 57, 85, 59, 70, 84, 43, 122]

In terms of the stylometry, the features looked upon would remain the function words until 1994 with the introduction of character n-grams in the authorship problem by Kjell [56]. The idea consists in capturing lexical preferences without any linguistic background, such as knowing language function words. Also, Forstall and Scheirer [32] also argue that character n-grams capture the sound of the words, being a valuable feature in authorship atribution of poetry.

Shortly after the first authors started using character n-grams, word n-grams would start being used in 2002 [44]. The idea behind it is that some authors might

have a preference for some expressions composed of two or more words in sequence, therefore captured by n-grams of these words.

Those features seem to follow the evolution in processing power and techniques emerging in computer science, moving from univariate to multivariate analysis and even expanding from simple function words, usually a set of a hundred of them, to character n-grams and word n-grams. In authorship analysis works, expanding the feature space, within lexical features, improves the classification but also requires additional processing power or better classifiers to deal with the amount of features generated.

Following the trends in machine learning, the field would turn to Suport Vector Machines (SVM) in the early 2,000s through the works of de Vel et al. [28] over e-mails and Diederich et al. [29] in German newspapers. This happened since SVM's are able to deal with a larger number of features and also they do not require any customization. The success of SVM in classifying authors would spread over the works in authorship attribution [60, 1, 63, 5, 37, 94, 2, 111, 32].

The robustness of these approaches with n-grams and SVM are further discussed in a recent report from Stamatatos [113] who has further investigated the question of whether the n-grams remain robust within cross-topic authorship attribution. Tables 2.2, 2.3, 2.4 and 2.5 list some of the works and their respective feature sets and classifiers.

Table 2.2: Works in modern authorship attribution - Part I.

Source	**Features Used**	**Classifier**	**Corpus**
Mosteller and Wallace 1964 [93]	Small set of function words[5]	Naïve-bayes	Federalist papers
Morton 1978 [89]	Word Positions	χ^2 tests	British authors
Merriam 1979 [80], 1980 [81], 1982 [82]	Word Positions	χ^2 tests	Shakespeare and Thomas More
Burrows 1987 [15], 1989 [15], 1992 [18, 17]	Small set of function words	Multivariate analysis over PCA	British writers
Morton and Michaelson 1990 [92], Morton 1991 [91]	Word positions	CUSUM	Various authors
Matthews and Merriam 1994 [76]	Small set of function words	Neural networks	Shakespeare
Kjell 1994 [58]	Character n-grams	Neural networks	Federalist papers
Merriam and Matthews 1994 [85]	Function words	Neural networks	Shakespeare books
Ledger and Merriam 1994 [68]	Character n-grams	Multivariate analysis	
Bissel 1995 [13]	Word count	WCUSUM	British authors
Holmes and Forsyth 1995 [39]	Function words	Genetic algorithms	Federalist papers
Kjell et al. 1995 [59]	Character n-grams	K-nearest neighbours	Wall Street Journal
Lowe and Matthews 1995 [70]	Function words	Neural networks	Shakespeare
Martindale and McKenzie 1995 [74]	Words	Linear discriminant analysis and Neural Networks	Federalist papers
Mealand 1995 [78]	Function words	Multivariate analysis	Biblical texts
Baayen et al. 1996 [8]	Syntax	Neural network	Federalist papers

[5]Function words also called stop words are in general very frequent words with no semantical meaning. Appendix A shows examples of those words.

Table 2.3: Works in modern authorship attribution - Part II.

Source	**Features Used**	**Classifier**	**Corpus**
Merriam 1996 [84] Function words	Multivariate analysis over PCA	Shakespeare	
Tweedie et al. 1996 [118]	Function words	Neural Networks	Federalist papers
Argamon-Engelson et al. [6]	Function words and part-of-speech n-grams	Newspapers	
Tweedie and Baayen 1998 [117]	Function words	Multivariate analysis over PCA	English prose
Somers 1998 [109]	Word positions	WCUSUM	British writers
Binongo and Smith 1999 [11]	Function words	Multivariate analysis over PCA	Shakespeare
Hoorn et al. 1999 [43]	Character n-grams	Neural Networks, k-nearest neighbours and naïve-bayes	Dutch poets
Stamatatos et al. 2000 [114], 2001 [115]	Syntactic chunks	-	Greek newspapers
Waugh et al. 2000 [122]	Words	Neural networks	Federalist papers
Kukushkina et al. 2001 [65]	Character n-grams	Markov models	Russian texts
Chaski 2001 [23]	Syntax and punctuation	-	Women essays
Holmes et al. 2001 [40, 41]	Function words	Multivariate analysis over PCA	Letters ad articles
Baayen et al. 2002 [7]	Function words	Syntax multivariate analysis over PCA	Dutch texts
Benedetto et al. 2002 [9]	Character n-grams	-	Italian texts
Burrows 2002 [20, 19]	Function words	Multivariate analysis over PCA	Restoration-era poets
Hoover 2002 [44], 2003 [46, 47, 45]	Word n-grams	Multivariate analysis	Novels and articles
Khmelev and Tweedie 2002 [54]	Character n-grams	Markov models	Novels and articles
Binongo 2003 [10]	Function words	Multivariate analysis over PCA	Lewis Carroll
Clement and Sharp 2003 [25]	Character n-grams	Naïve-bayes	Movie reviews

Table 2.4: Works in modern authorship attribution - Part III.

Source	**Features Used**	**Classifier**	**Corpus**
Diederich et al. 2003 [29]	Words	SVM	German newspapers
Keselj et al. 2003 [53]	Character n-grams	Multivariate analysis	English novels, Greek newspapers
Khmelev and Teahan 2003 [55]	Character n-grams	Markov models	Russian texts
Argamon et al. 2003 [3]	Function words and part-of-speech n-grams	Winnow	British national corpus
Somers and Tweedie 2003 [110]	Word positions	WCUSUM	British writers
Peng et al. 2003 [95]	Word N-grams	Markov models	Federalist papers
Hoover 2004 [49, 48]	Words	Multivariate analysis over PCA	Novels and articles
Peng et al. 2004 [96]	Character and word n-grams	Naïve-bayes	Greek texts
van Halteren 2004 [119]	Word n-grams and syntax	Multivariate analysis	Dutch texts
Abbasi and Chen 2005 [1]	Characters, words, vocabulary	SVM	Arabic forum posts
Chaski 2005 [24]	Character and word n-grams	Linear discriminant analysis	Anonymous authors
Juola and Baayen 2005 [51]	Function words	Cross-entropy	Dutch texts
Zhao and Zobel 2005 [126]	Function Words	Naïve-bayes and k-nearest neighbours	Newswire stories
Koppel et al. 2005 [63]	Function words and part-of-speech n-grams	SVM	English students
Koppel et al. 2006 [62]	Function words and part-of-speech n-grams	BalancedWinnow	British national corpus
Zhao et al. 2006 [128]	Function words and punctuation	SVM	Associated press stories and English novels
Madigan et al. 2006 [71]	Characters and words	Bayesian regression	Federalist Papers
Zheng et al. 2006 [129]	Characters, function words and syntax	Neural networks and SVM	English and Chinese newsgroups

Table 2.5: Works in modern authorship attribution - Part IV.

Source	**Features Used**	**Classifier**	**Corpus**
Li et al. 2006 [69]	Characters, function words and syntax	Neural networks and SVM	English and Chinese newsgroups
Argamon et al. 2007 [5]	Function words and syntax	SVM	Novels and articles
Burrows 2007 [21]	Words	Multivariate analysis	Restoration poets
Pavelec et al. 2007 [94]	Conjunction types	SVM	Portuguese newspapers
Zhao and Zobel 2007 [127]	Function words and part-of-speech n-grams	Information gain	British authors
Stamatatos 2008 [111]	Character n-grams	SVM	English and Arabic news
Argamon et al. 2009 [4]	Words and systemic functional linguistics	Bayesian regression	Blogs and student essays
Forstall and Scheirer 2009 [32]	Character n-grams	SVM	Shakespeare

2.3 Authorship Attribution on Micro-Messages and short texts

Nowadays, we face a new challenge for authorship attribution due to the growth of social media, specifically with micro-blogs. They are composed of very short messages, which makes the analysis substantially harder. Although some authors have worked with short messages and texts [28, 60, 101, 37, 2], only a few have tackled the authorship attribution problem in micro-blogs thus far [67, 14, 105, 86, 103].

Current literature shows that the methods relying upon the SVM classifier [105, 86, 103] outperform other approaches to authorship attribution on micro-blogs [103], namely Naïve-Bayes [14] and Source Code Authorship Profiling (SCAP) [67]; almost all approaches use the same set of features, character-level and word-level n-grams [103]. These features normally follow standard practice on web data: $n = 4$ for character n-grams, and $n = (2, \ldots, 5)$ for word n-grams in a traditional bag-of-words model. This limitation was up to now mostly regarded to the impossibility of carving useful decision spaces with traditional implementations of the used classifiers. Table 2.6 shows the authors who tackled the problem and the database used in the analysis.

Table 2.6: Works in short text authorship attribution.

Source	**Features Used**	**Classifier**	**Database**
de Vel et al. 2001 [28]	Various	SVM	E-mails
Koppel and Schler 2003 [60]	Function words, part-of-speech n-grams	SVM	E-mails
Sanderson and Guenter 2006 [101]	Word sequences	Markov chains	Short essays[6]
Hirst and Feiguina 2007 [37]	Syntatic bigrams	SVM	Short essays[7]
Abbasi and Chen 2008 [2]	Characters, function words and syntax	SVM and PCA	Online content
Layton et al. 2010 [67]	Word and character n-grams	SCAP	Micro blog texts
Boutwell 2011 [14]	N-grams and cellphone data[8]	Naïve-bayes	Micro blog texts
Silva et al. 2011 [105]	Internet Metalanguage	SVM	Micro blog texts
Layton et al. 2012 [66]	Character n-grams	SCAP	IRC messages
Mikros et al. 2013 [86]	Greek specific features	SVM	Micro blog textsr[9]
Koppel et al. 2013 [103]	Word and character n-grams	SVM	Micro blog texts

[6]Texts up to 1,850 characters.
[7]Texts up to 500 words.
[8]This work was mainly focused on identifying the user cellphone along with the lexical features.
[9]They joined more than one text into a single document.

Chapter 3

Proposed Method

In this work, we propose a methodology for authorship attribution on micro-blog texts. The innovative aspect of our approach is the use of a complete set of features relying on patterns extracted from unigrams to 5-grams of words with 4-grams of characters in a bag-of-words model dynamically created for each user.

Previous work did not use unigrams as they are allegedly captured by the character n-grams, and due to the explosion in the feature representation harming the classification process. Differently, we show that unigrams, with the character n-grams, substantially improve the classification accuracy.

We also deal with the high-dimensional data representation using an improved version of the SVM classifier for large-scale image classification [123]. Although there are many fast linear solvers for SVM, such as SVM LibLinear[1], they perform as the traditional SVM, just being a little faster.

3.1 Baseline System for Authorship Attribution

Following the idea of authorship attribution, we have defined a general guideline for an authorship attribution system, as shown in Fig 3.1, starting with an unknown message and elaborating a list of suspects. In this work, we propose an approach that scales well with a large amount of suspects,

We proceed by extracting old data from the suspects accounts (Sec. 3.2). This data is preprocessed (Sec. 3.3) to remove very sparse features, too short and non-

[1] www.csie.ntu.edu.tw/~cjlin/liblinear/

English messages. All of the preprocessed texts are then used to create a dictionary, a bag-of-words model (Sec. 3.4), of character and word n-grams (Sec. 3.5 and Sec. 3.6).

The messages are processed again, based on the dictionary, generating the feature vectors both for the extracted data used in training and testing the classifier as well as the message whose authorship we wish to discover. The training vectors are fed to a large-scale SVM classifier (Sec. 3.7), and tested against the initial message. The result will then point out to the most probable suspect.

Although the classifier points out to a single class, in our experiments, we discuss the use of the output function from the classifier as a mean to build a cumulative matching, used to prioritize the most likely suspects.

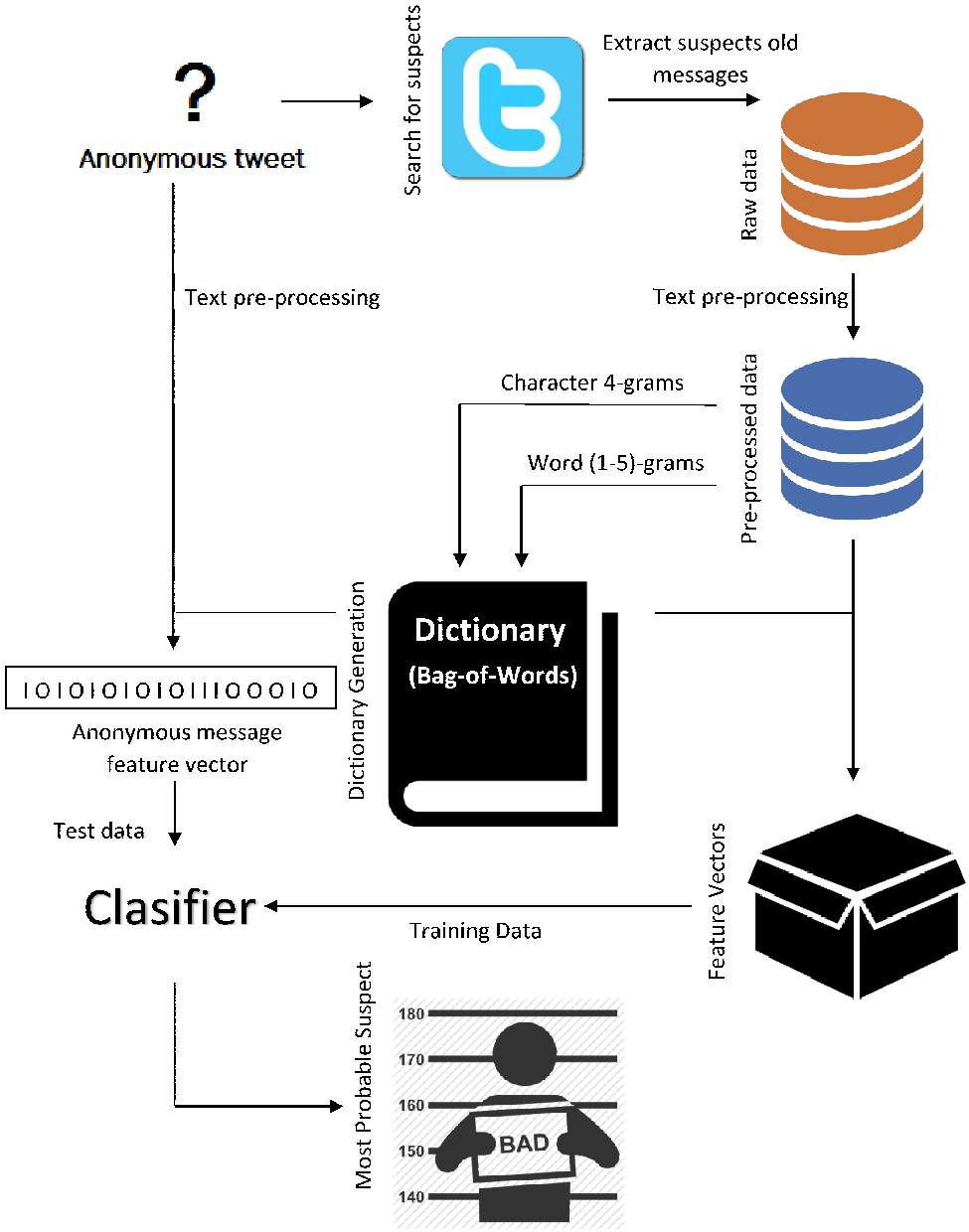


Figure 3.1: General pipeline for micro-blog authorship attribution.

3.2 Data Extraction

We could not use data from the literature due to the Twitter data policy. Searching Twitter for English function words present in Appendix A, then we got results from english speaking public users[2]. These are used to build a list of public users from whom we can extract all the tweets from these users with the Twitter API[3].

Since the goal is authorship attribution, we removed all the retweets, which are messages retransmitted from other users. This is done by removing all tweets marked by the API with a specific flag, and also tweets containing the metatag RT [67, 103]. We also removed non-english tweets using python library guess-language [97] which uses the spell-checking library pyenchant [52] to build an accurate prediction of the language of texts with three or more words. For this reason, we removed all tweets that had only one or two words.

3.3 Text Preprocessing

For authorship attribution, little preprocessing is required. This is due to the way people write: preferences for letter capitalization, word suffixes and even grammatical mistakes are part of their writing style. It makes no sense to stem words or even grammar checking them. This would greatly reduce the features to be analyzed, but it would also remove many idiosyncrasies, which are unique to a user [67, 103] such as internet expressions, common grammar mistakes and abbreviations.

Therefore, preprocessing was focused on grouping very sparse characteristics such as numbers, dates, times and URLs. Those are relevant characteristics, but it is unlikely that a user will be sharing, for example, the same date many times. So we took away the original text, replacing it with a standard tag that represents each of the contents replaced.

Moreover, it is proven that hashtags[4] and user references[5] make authorship attribution easier [67]. A user might repeatedly use the same hashtag or talk too much

[2]Public user data is not subject to any form of protection from Copyright law. Twitter data extraction policy still applies, therefore this data can be used but not shared.

[3]The Twitter API only allows the extraction of the last 3200 tweets from each user.

[4]Hashtags are keywords used in social media to make searching for a subject easier; usually they are preceded by a # character and found amidst the text.

[5]The users of social media can send messages to other users using an @ followed by their nickname.

to a single person. It also makes the method unreliable, because a user might make references to the same person across his messages, creating a bias in this particular feature; and any message with a reference to the same person would be misclassified as being from that user.

In the following example we show three tweets before and after preprocessing.

Before preprocessing:

Tweet 1: "Do not forget my bday is on 27/03 #iwantgifts"

Tweet 2: "@maria I will be sleeping @00:00AM"

Tweet 3: "Check out this amazing website: www.ic.unicamp.br"

After preprocessing:

Tweet 1: "Do not forget my bday is on DAT TAG"

Tweet 2: "REF I will be sleeping @TIM"

Tweet 3: "Check out this amazing website: URL"

These small changes in the messages not only reduce the sparsity of the features, but also make the process more robust. In doing so, we avoid that some particular user data, such as a link to his personal website for example, create a bias in the model by being repeated many times.

3.4 Bag-of-Words Model

The bag-of-words is a traditional model in natural language processing. It is an orderless document representation of feature frequencies from a dictionary [100]. Although the term word is used, the dictionary may consist of groups of n words (n-grams), groups of letters or other textual features that can be extracted. For many tasks such as information retrieval and sentiment analysis, it has been proven that function words should be removed [72]. For other uses, which is the case of this work, function words can provide more information about the author than more meaningful words [32].

These results are derived from Zipf's law [130], which states that the frequencies of the words are inversely proportional to their rank on the frequency table. Since the most used words vary for each user, but they are always amongst the function

words, we can derive many of the nuances of each user by looking at the function words alone [79]. In this work, we have to use more features, due to the shortness of the messages and, therefore, the words contained in a single document are likely to appear only once. Fig. 3.2 shows the variation of the most used words among 10 authors.

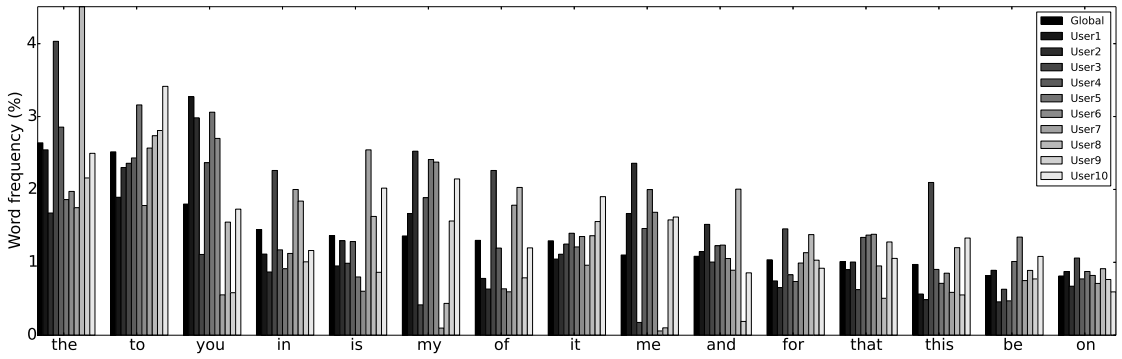


Figure 3.2: Most used word frequencies for 10 different Twitter authors and a global frequency obtained joining the frequencies of all ten authors.

The bag-of-words model can be exemplified with the following two excerpts of Shakespeare:

Text 1: "To be, or not to be, that is the question."

Text 2: "To die, to sleep. To sleep, perchance to a dream;"

We then create a dictionary that maps each feature, in this example the words, to a number as shown in Table 3.1, then we can use these numbers to create the feature vectors later.

Table 3.1: Bag-of-words example.

Number	1	2	3	4	5	6	7	8	9	10	11	12
Word	to	be	or	not	that	is	the	question	die	sleep	perchance	dream

With this dictionary, we can create feature vectors based on the frequency of the words. In this example, we use the raw frequency (simply the count of words), but other operations might be considered as well, such as tf-idf [32]. In our work, we used binary vectors that indicate the occurrence of the word (1) or its absence (0):

Feature vector 1: [2, 2, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0]

Feature vector 2: [4, 0, 0, 0, 0, 0, 0, 0, 1, 2, 1, 1]

This model is used throughout the work and the dictionaries are constructed using word and character n-grams as presented in Sec. 3.5 and Sec. 3.6 .

3.5 Character n-grams

Character n-grams are often used for authorship attribution on web-based texts as they capture unusual features, such as emoticons and special use of punctuation. They help mending the effect of small typos that authors do not repeat very often. For example, with the word "misspeling", the generated character 4-grams would still have "miss", "issp", "sspe", "spel" and "ling", in common with the 4-grams generated for the correct word "misspelling".

Following the literature [67, 101, 103], herein we focus on character 4-grams as white spaces and metatags are included in the n-grams. White spaces are appended at the beginning and at the end of each tweet. Also we discard any character 4-grams which does not appear at least twice for the same author [103]. This approach makes the SVM run much faster and removes noisy features that are unlikely to appear again within the same user.

The features used are case-sensitive, since the author's preference for capitalization of letters is also one of the traits that can be used for identification. That is because many users on social media have a preference of capitalizing some words to emphasize them or even writing in the infamous camel-case[6].

The following sample shows how we compute the 4-grams from a simple phrase:

Text: "2B!!!! or n2B!!!!!!! ;)"

This would yield the following 4-grams:

" 2B!"; "2B!!"; "B!!!"; "!!!!"; "!!! "; "!! o"; "! or"; " or "; "or n"; "r n2"; " n2B"; "n2B!"; "2B!!"; "B!!!"; "!!!!"; "!!!!"; "!!!!"; "!!!!"; "!!! "; "!! ;"; "! ;)"; " ;) "

[6]Camel-case is a style that alternates between lower and upper case letters e.g.: "ExAmPlE"

This approach is able to isolate the repeating pattern of exclamation marks and the emoticon[7], which are inherently of the social media language. If we included character n-grams for other n, we would generate redundant data either by repeating something captured by $n = 4$, when selecting $n < 4$, or by capturing a dirty version of word unigrams, with $n > 4$. Having the same feature duplicated on the database makes it stronger than it should be, thus creating a bias towards it, what is undesirable.

3.6 Word N-grams

The use of the traditional bag-of-words model is proven to be useful for authorship attribution of micro-messages [103]. Because tweets are very small, it is commonplace for users to repeat short phrases and the same words all over their micro-messages. We also include punctuation sequences to find the n-grams, considering that these sequences might be part of the phrases.

We used n-grams for words with $n \in (1, \ldots, 5)$. According to the literature, character level n-grams should generally capture the unigrams (1-grams) and their use would increase the feature vectors substantially. Herein we show their complementarity regarding the normal characterization used in the literature. Special meta-tags were also considered at the beginning and at the end of each tweet to distinguish words frequently used to start and to end messages. These features are also case-sensitive.

The use of word n-grams is as follows[8]:

> **Text:** "To be, or not to be, that is the question."
>
> **Unigrams:** ("To", "be", "or", "not", "to", "be", "that", "is", "the", "question")
>
> **Bigrams:** ("BEGIN To", "To be", "be or", "or not", "not to", "to be", "be that", "that is", "is the", "the question")

[7]The emoticons were not replaced by metatags because they are part of a user language in the internet. A user will have a particular combination of emoticons and patterns they use.[105]

[8]In the example, we used "BEGIN" and "END" to mark the start and end of the text, although when implementing this we used non-printable characters to avoid mismatching with the respective English words.

3-grams: ("BEGIN To be", "To be or", "be or not", "or not to", "not to be", "to be that", "be that is", "that is the", "is the question")

...

When using all n-grams and $n = 4$ word n-grams, the proposed method has feature vectors varying according to the number of training examples and analyzed users. For instance, the feature dimensionality varies from 20,000-d vectors (50 users and 50 training tweets per user) to around 500,000-d vectors (500 users and 500 tweets per user). Although the traditional literature for authorship attribution has used $n < 4$ as default (e.g., the authors in [32] used $n = 2$), for micro messages, we need a larger n to capture internet meta-language properties such as emoticons, onomatopoeia, abbreviations, and others.

3.7 Large-Scale Classification

While most authors use a traditional formulation of the SVM classifier, we observed in our experiments that it does not handle large data and high-dimensional feature vectors very well. This is indeed one of the reasons researchers have avoided unigrams for feature representation.

Differently, we rely upon Power Mean SVM kernel (PMSVM) formulation [123], which was originally proposed for large-scale image classification. The power mean kernel generalizes many kernels in the additive kernel family. These naturally arise in applications such as image and text classification, where the data is well represented by histograms or bag-of-word models. Also, this kernel family is not very sensitive to parametrization, avoiding overfitting to the training data. It is shown that additive kernels are more accurate in most vision problems [123].

The power mean kernel aggregates the advantages of linear SVM and non-linear additive kernel SVM. It performs faster than other additive kernels because rather than approximating the kernel function and the feature mapping, it approximates the gradient function using polynomial regression. This approach outperforms fast linear SVM solvers (e.g., LibLinear SVM, and Coordinate Descent SVM) in about 5×, and also, in the state-of-the-art additive kernel SVM training methods in about 2× (e.g., HIK SVM) [123]. Therefore, this kernel converges using only a small fraction of the iterations needed for the linear classification when dealing with a large number of features and big data [123].

An SVM kernel κ is additive if it can be written as a sum of a scalar function for each feature dimension d, i.e., for two vectors x and y,

$$\kappa(x, y) = \sum_{i=1}^{d} \kappa(x_i, y_i) \tag{3.1}$$

and a power mean function M_p is defined by a real number $p \in \mathbb{R}$ and a set of positive numbers $x_1, \ldots, x_n \in \mathbb{R}$:

$$M_p(x_1, \ldots, x_n) = \left(\frac{\sum_{i=1}^{n} x_i^p}{n} \right)^{\frac{1}{p}}. \tag{3.2}$$

Many of the additive kernels are special cases of the power mean function, such as the χ^2, Histogram Intersection and the Hellinger's kernels [123]. The power mean kernel for two vectors $x, y \in \mathbb{R}_+$ proposes an approach that generalizes those three kernels:

$$M_p(x, y) = \sum_{i=1}^{d} M_p(x_i, y_i). \tag{3.3}$$

It is proven in [123] that the kernel is well defined for any value of p. Usually, this formulation would lead to higher training times, but the authors of PMSVM use the coordinate descent method with a gradient approximation to solve the dual SVM problem. As a result, training is also faster and the approximation avoids the overfitting to the training data [123].

Chapter 4

Experimental Results and Validation

Given the recent changes in the Twitter data policy, which forbids data exchange, we could not test our approach directly on the same dataset used by other authors. Therefore, to make a fair comparison between our methods and the state of the art, we reproduced the experiment as described in Koppel et al. [103] and implemented their best method. The authors proposed the current state-of-the-art method, and performed an extensive comparison with other works for micro-blog authorship attribution [67, 14], outperforming them all.

4.1 Dataset and Preprocessing

The dataset comprises 10^7 tweets from 10^4 writers in English. Each tweet is at most 140-character long and may include hashtags, user indications and links. The dataset was collected across several days, in march of 2014, within the latest 3,200 tweets from each user. To restrict the search to English speaking users, we searched for English function words in the API [67].

Preprocessing of each micro-message includes removing all non-english tweets, tweets with less than three words and retweets, those marked as retweets or any tweet containing the meta-tag RT. For sparsity reasons, we replaced numbers, URLs, dates and timestamps by the metatags NUM, URL, DAT, TIM, respectively. Moreover, the hashtags and user references were replaced, since they enrich the feature set for

author attribution making it an easier but unreliable task [67].

4.2 Independent Features Usefulness

In order to validate our hypothesis about the features used, we performed independent tests with each set of features, namely: word n-grams for $n = 1, \ldots, 5$ and character 4-grams.

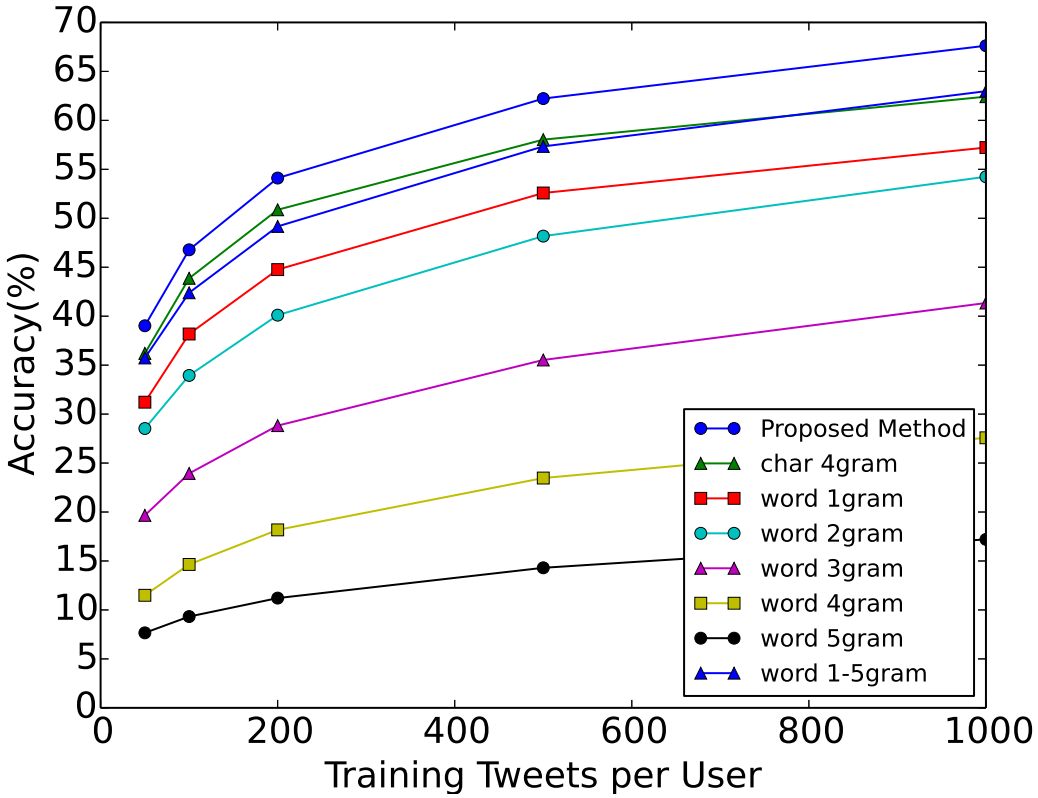


Figure 4.1: Relevance of each feature used alone and joined features in the proposed method considering 50 users.

In Fig. 4.1, it is possible to see that even the more sparse feature – which is word 5-grams – still fares above the random baseline for 50 users (2% accuracy). It is shown that the most relevant independent feature is the character 4-grams. This

can be explained because it captures the internet metalanguange such as repetitions of punctuations and emoticons, amongst other stylometry [103].

The figure also shows that unigrams are a highly relevant feature. When combined with the other word n-grams, they present an accuracy similar to the character 4-grams. This reflects the users' preferences for some words they are more comfortable with and which are not captured by character n-grams due to their size or likelyhood with other word prefixes and suffixes. Joined with other n-grams, it captures all the author preferences, from words to small characteristic sentences and expressions.

4.3 Feature Set

Although unigrams have not been used previously for micro-blog authorship attribution for the reasons discussed before, our findings show that they offer a substantial contribution to the problem. The use of unigrams generates a greater number of features to be used. Since our solution relies upon PMSVM, the number of features has less impact on the classification time, improving accuracy significantly.

Using unigrams for feature representation, our solution outperforms the literature by a margin of over 10% for 50 users and 500 tweets per user, as Fig. 4.4 depicts. Most importantly, the method proposed in the literature did not converge for all tests in more than two days of computing when we used 1,000 training tweets per user. This is a significant margin explained by the use of a classifier more adequate for big data and large number of features. Note the steady improvement in the classification task was due to the use of unigrams as features.

We also evaluated how our approach deals with an increasing number of users, testing our solution with 500 users. Figs. 4.3 and 4.5 show the results for these experiments. We can see that the classifier nicely handles hundreds of users and continues to perfect as more training messages are used per user.

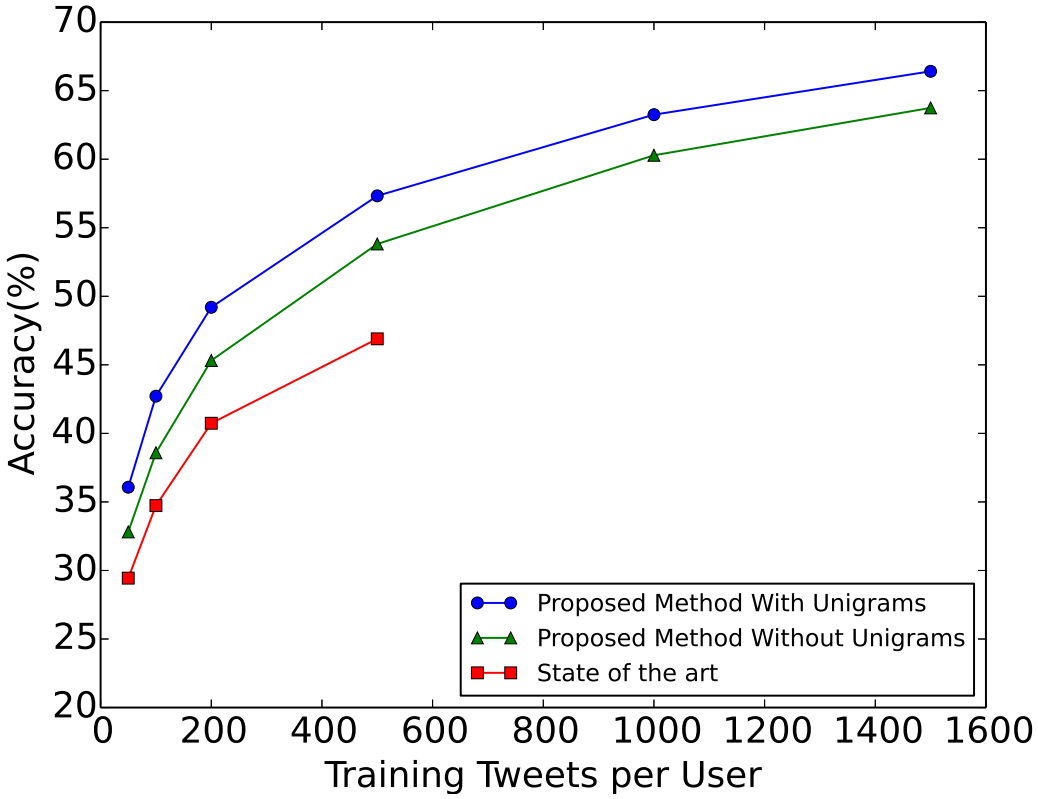


Figure 4.2: Classification accuracy comparison between our aproaches *vs*. the state-of-the-art [103] for 50 users.

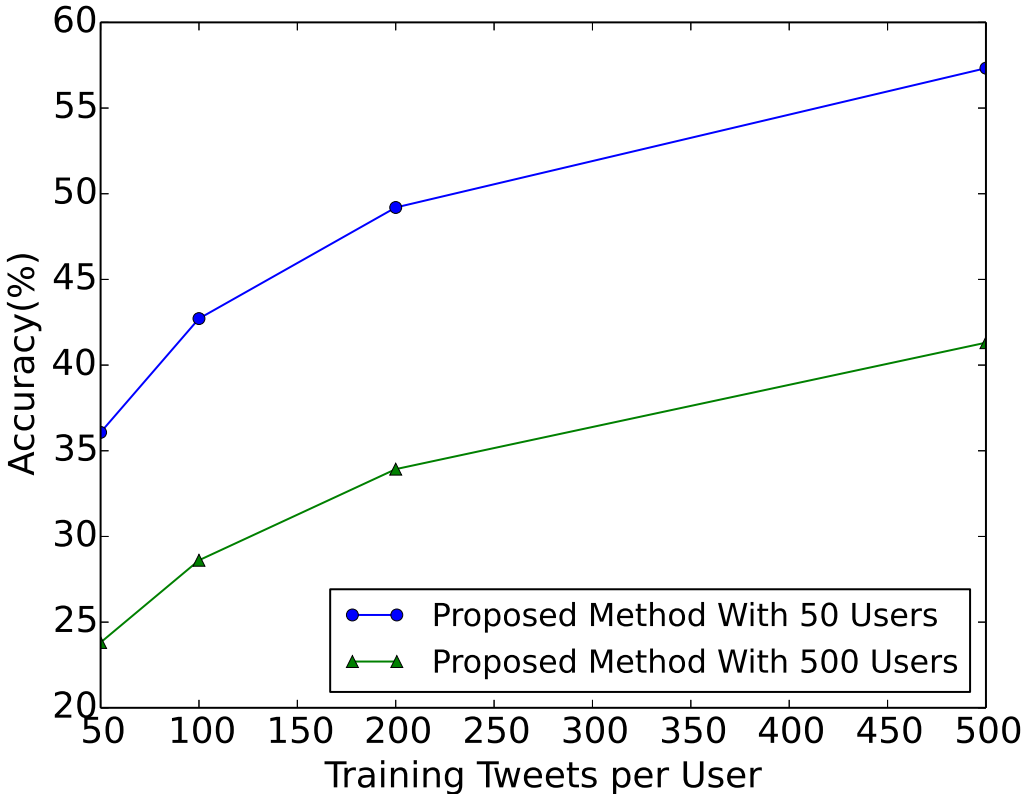


Figure 4.3: Classification accuracy of the proposed method for 50 and 500 users.

4.4 SVM × PMSVM

Here we explore how the proposed method compares to others with respect to the computational efficiency considering 50 users and a varying number of training messages per user. Fig. 4.4 shows that the proposed classification technique runs between $100\times$ to $1,000\times$ faster than [103], which relies on traditional SVM formulation. Our solution based on PMSVM also outperforms [103] in terms of classification effectiveness. Using 50 tweets per user for training the method is 2.5% more effective; when using 500 training tweets per user, the difference between the methods increases to more than 4% as Fig. depicts. This difference tends to increase as more training data is used since the proposed solution nicely handles large-scale data. Note again that PMSVM is neither sensitive to the parameter C nor to other parameters of the SVM classifier, so there is no need for a parameter grid search for fine tuning [123].

For this analysis, we use a 2.30GHz 3rd generation Intel Core i7-3610QM computer with 8GB 1600MHz DDR3 SDRAM (2 DIMM) running Fedora Linux. For better visualisation, we used a log scale. Both curves approach a quadratic function of the number of training data.

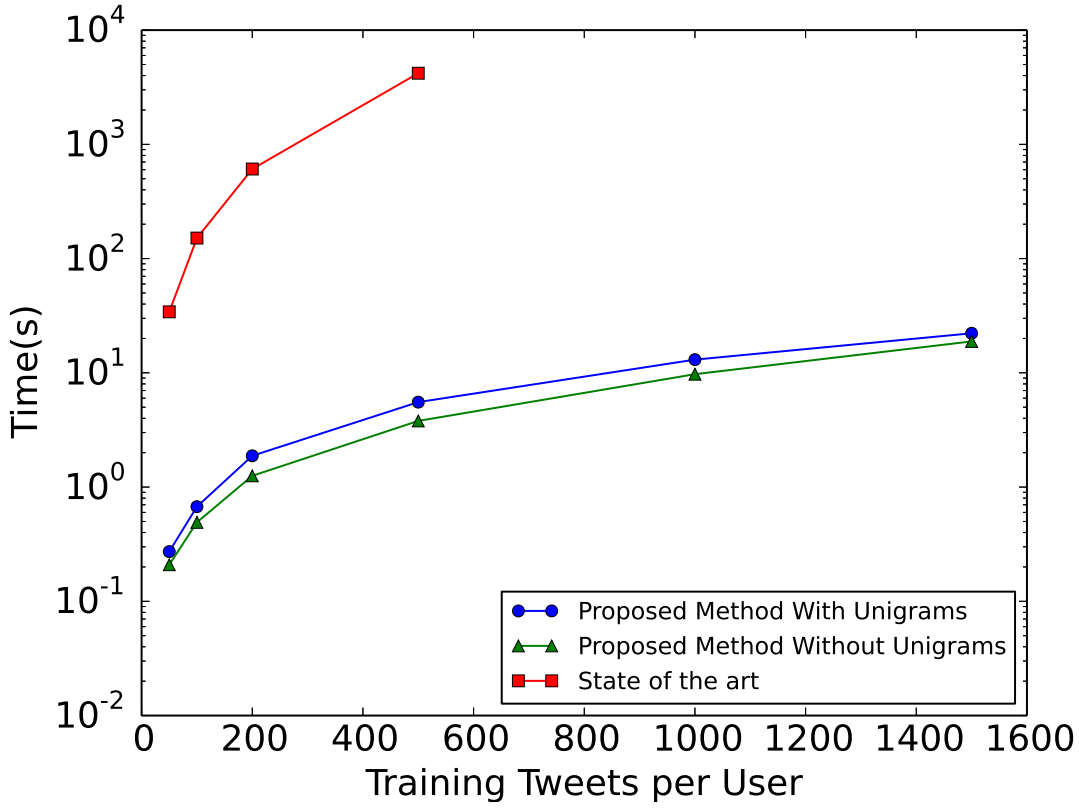


Figure 4.4: Eficiency comparison between the training time for our aproaches *vs*. the state-of-the-art [103] for 50 users

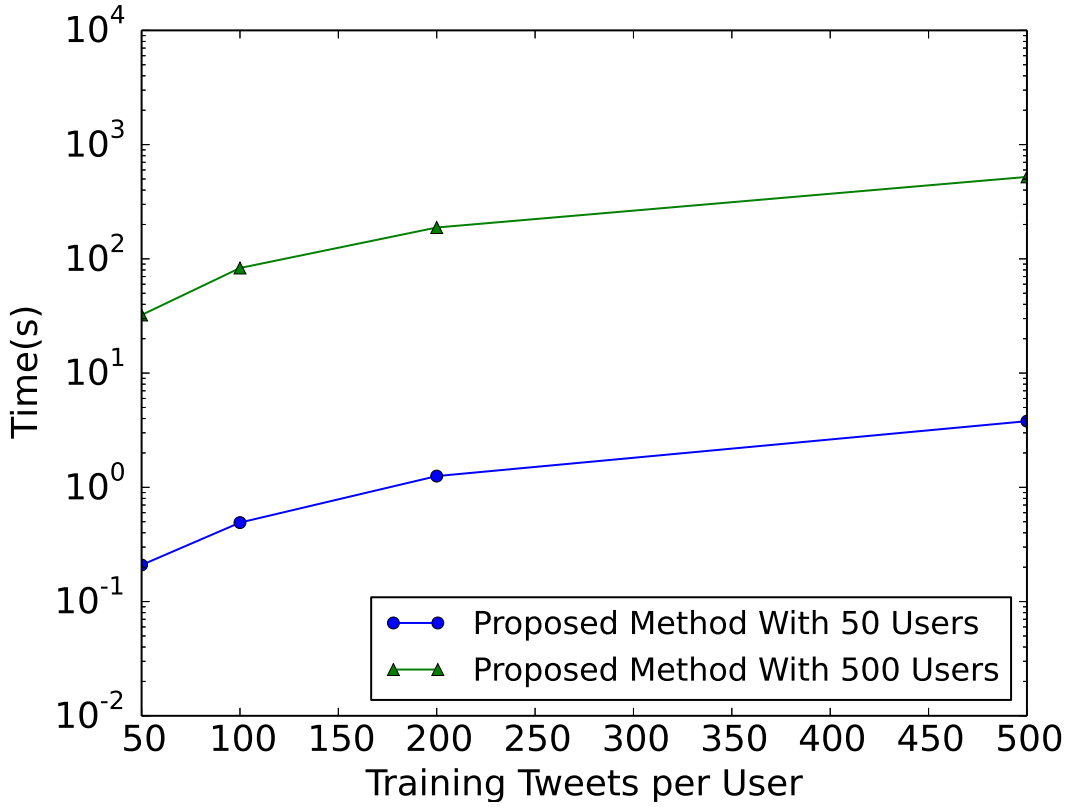


Figure 4.5: Training time consumption of the proposed method for 50 and 500 users.

4.5 Search-Space Reduction

Although traditional authorship analysis intends to find the author of a given text, that might not be the case with micro-web messages. The very nature of short messages make it hard to attribute them to a single author. Also, the problems involved in the task, such as hoaxes, impersonations and identity stealing, make it an open problem whereby the user may not be among the suspects, and the classifiers will always point out to someone. This suggests that the effort should also be on reducing and prioritizing the suspects rather than pointing to a single culprit.

Instead of seeking for the most probable user, we can rank the classes according to the output function, and then show how well we can reduce the search space of the problem. We tested the method with 500 users and a varying number of tweets per user. The Cumulative Matching Curve (CMC) shows the accuracy of finding the author of a tweet considering the top N users.

The classifier starts with a classification accuracy of 35% when using 50 tweets per user for training. Considering the random baseline of 0.2%, this is a remarkable result for micro-blog authorship attribution. In more than 65% of the cases, the correct user will be among the top 50 of 500 users when we use 200 training tweets per user (see Fig. 4.6). In a real scenario, this would reduce the number of suspects to 10% of the original size in more than half of the scenarios.

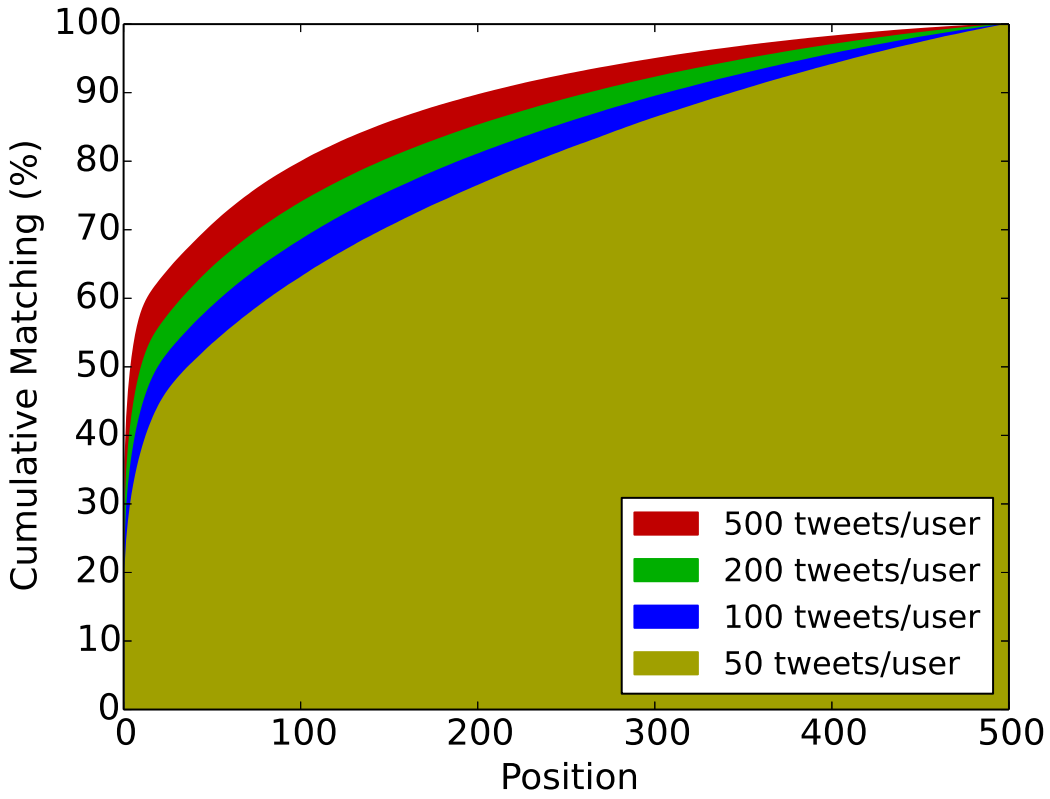


Figure 4.6: Cumulative Matching Curve for 500 users with a varying number of training tweets per user.

Chapter 5

Conclusion and Future Work

This dissertation showed that the information explosion and scarcity of information per user in micro-blog social networks require better methods to cope with the problem of authorship attribution. Even when dealing with micro-messages, the problem grows really fast, because of the large number of users and messages involved. Also, there is a huge amount of features to be considered, specially due to the use of unconventional punctuation, abbreviations and internet meta-language. This generates feature vectors with high dimensions, which in turn can lead to the curse of dimensionality [12]. Therefore, we need to use classifiers that better handle large datasets, as well as the high number of dimensions generated by the problem.

Although the accuracy of the classification does not hit perfection, the method discussed herein is surely an advance when compared to the current state-of-the-art methods and opens the possibility for other authors to explore features and enhanced classifiers overseen thus far. In addition, the cumulative matching approach analysis shows that our method can greatly reduce the number of users to be analyzed in a real situation. We also showed that some features, which capture great stylistic patterns, not used before due to technical limitations, now can be considered because of their discrimination power.

5.1 Real World Application

As discussed in Chapter 2, there was much controversy on the use of authorship attribution in trials. However, that does not exclude the pratical application of

authorship attribution. The method proposed here can be used on investigations, especially for search-space reduction.

The results are very solid, reaching 75% accuracy for 50 authors, and may indeed be used while looking for clues in a real scenario. For further applications, we should also look deeper at the structure of a social network. Such work might help even to find criminal groups using social media via the social graph.

5.2 Big Data

The amount of data present on social media is growing faster each day. In this work, we looked uppon a more robust classifier to deal with such amount of data. Figs. 4.4 and 4.4 show that even with 30 times the amount of training data per user, we were still able to perform better than the state-of-the-art classifier.

The curves on those graphs also indicate that by using more training data we can get better results. Since PMSVM is able to deal with large amounts of data, while being more accurate, the proposed method represents a great improvement over the state-of-the-art methods given the big data nature of the problem.

Big data can be classified over at least three aspects, which are: volume, speed and variety of data. Volume and speed of data on social media is clearly large. In this work, we only dealt with textual features, but nonetheless the social media data is much more variable than just text. Future work should also look at pictures and videos posted online to further improve authorship attribution on social media.

5.3 Social Media Native Features

Our method takes into account some features native of social media, outside the realm of regular texts, such as hashtags, weblinks and user messages. However, there are still many other native features to be explored in social media, including the aforementioned social graph. Also other features such as the message times, device used, browser used, amongst other details, could be explored.

5.4 Denser Features

Since we use dynamic features extracted from raw text, and tweets contain few words, the features used are very sparse in nature. In the course of our work, we tried many classical techniques to reduce the sparsity of our feature vectors such as PCA and random selection of features. Random selection always performed worse no matter the size of the features extracted, and PCA failed to run due to the overwhelming number of examples and features.

Future work could explore other sets of features and dimensionality reduction techniques that might improve the classification task being present in almost all tweets and, therefore, avoiding the currently very sparse feature vectors.

Bibliography

[1] Ahmed Abbasi and Hsinchun Chen. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, 20(5):67–75, 2005.

[2] Ahmed Abbasi and Hsinchun Chen. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)*, 26(2):7, 2008.

[3] Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. Gender, genre, and writing style in formal written texts. *Journal of Writing and Writing Courses*, 23:3, 2003.

[4] Shlomo Argamon, Moshe Koppel, James W Pennebaker, and Jonathan Schler. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119–123, 2009.

[5] Shlomo Argamon, Casey Whitelaw, Paul Chase, Sobhan Raj Hota, Navendu Garg, and Shlomo Levitan. Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, 58(6):802–822, 2007.

[6] Shlomo Argamon-Engelson, Moshe Koppel, and Galit Avneri. Style-based text categorization: What newspaper am i reading. In *AAAI Workshop on Text Categorization*, pages 1–4, 1998.

[7] Harald Baayen, Hans van Halteren, Anneke Neijt, and Fiona Tweedie. An experiment in authorship attribution. In *Journal of Textual Data Statistical Analysis*, pages 29–37. Citeseer, 2002.

[8] Harald Baayen, Hans Van Halteren, and Fiona Tweedie. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3):121–132, 1996.

[9] Dario Benedetto, Emanuele Caglioti, and Vittorio Loreto. Language trees and zipping. *Physical Review Letters*, 88(4):048702, 2002.

[10] José Nilo G Binongo. Who wrote the 15th book of oz? an application of multivariate analysis to authorship attribution. *Chance*, 16(2):9–17, 2003.

[11] José Nilo G Binongo and Michael W A Smith. The application of principal component analysis to stylometry. *Literary and Linguistic Computing*, 14(4):445–466, 1999.

[12] Christopher M Bishop. *Pattern recog. and machine learning*, volume 1. Springer, 2006.

[13] Alfred F Bissell. Weighted cumulative sums for text analysis using word counts. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pages 525–545, 1995.

[14] Sarah R Boutwell. Authorship attribution of short messages using multimodal features. Master's thesis, Naval Postgraduate School, Monterey, CA, USA, 2011.

[15] John F Burrows. Word-patterns and story-shapes: The statistical analysis of narrative style. *Literary and Linguistic Computing*, 2(2):61–70, 1987.

[16] John F Burrows. 'an ocean where each kind...': Statistical analysis and some major determinants of literary style. *Computers and the Humanities*, 23(4-5):309–321, 1989.

[17] John F Burrows. Computers and the study of literature. *Computers and Written Texts*, pages 167–204, 1992.

[18] John F Burrows. Not unles you ask nicely: The interpretative nexus between analysis and information. *Literary and Linguistic Computing*, 7(2):91–109, 1992.

[19] John F Burrows. The englishing of juvenal: computational stylistics and translated texts. *Style*, 36(4):677–699, 2002.

[20] John F Burrows. 'delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3):267–287, 2002.

[21] John F Burrows. All the way through: testing for authorship in different frequency strata. *Literary and Linguistic Computing*, 22(1):27–47, 2007.

[22] David Canter. An evaluation of the "cusum" stylistic analysis of confessions. *Expert Evidence*, 1(2):93–99, 1992.

[23] Carole E Chaski. Empirical evaluations of language-based author identification techniques. *Forensic Linguistics*, 8:1–65, 2001.

[24] Carole E Chaski. Who's at the keyboard? authorship attribution in digital evidence investigations. *International Journal of Digital Evidence*, 4(1):1–13, 2005.

[25] Ross Clement and David Sharp. Ngram and bayesian classification of documents for topic and authorship. *Literary and Linguistic Computing*, 18(4):423–447, 2003.

[26] Pieter de Haan and Erik Schils. The qsum plot exposed. In *International Computer Archive of Modern and Medieval English*, 1993.

[27] Sophia E De Morgan. *Memoir of Augustus de Morgan, by His Wife Sophia Elizabeth de Morgan, with Selections from His Letters*. Longmans, Green and Company, 1882.

[28] Olivier De Vel, Alison Anderson, Malcolm Corney, and George Mohay. Mining e-mail content for author identification forensics. *ACM Sigmod Record*, 30(4):55–64, 2001.

[29] Joachim Diederich, Jörg Kindermann, Edda Leopold, and Gerhard Paass. Authorship attribution with support vector machines. *Applied Intelligence*, 19(1-2):109–123, 2003.

[30] Peter Dixon and David Mannion. Goldsmith's periodical essays: a statistical analysis of eleven doubtful cases. *Literary and Linguistic Computing*, 8(1):1–19, 1993.

[31] Alvar Ellegård. *A Statistical method for determining authorship: the Junius Letters, 1769-1772*, volume 13. Göteborg: Acta Universitatis Gothoburgensis, 1962.

[32] Christopher W Forstall and Walter J Scheirer. Features from frequency: Authorship and stylistic analysis using repetitive sound. In *Annual Chicago Colloquium on Digital Humanities and Computer Science*, 2009.

[33] Wilhelm Fucks. On mathematical analysis of style. *Biometrika*, pages 122–129, 1952.

[34] RA Hardcastle. Cusum: A credible method for the determination of authorship? *Science & Justice*, 37(2):129–138, 1997.

[35] Robert A Hardcastle. Forensic linguistics: An assessment of the cusum method for the determination of authorship. *Journal of the Forensic Science Society*, 33(2):95–106, 1993.

[36] Michael L Hilton and David I Holmes. An assessment of cumulative sum charts for authorship attribution. *Literary and Linguistic Computing*, 8(2):73–80, 1993.

[37] Graeme Hirst and Ol'ga Feiguina. Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, 22(4):405–417, 2007.

[38] David I Holmes. The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, 13(3):111–117, 1998.

[39] David I Holmes and Richard S Forsyth. The federalist revisited: New directions in authorship attribution. *Literary and Linguistic Computing*, 10(2):111–127, 1995.

[40] David I Holmes, Lesley J Gordon, and Christine Wilson. A widow and her soldier: Stylometry and the american civil war. *Literary and Linguistic Computing*, 16(4):403–420, 2001.

[41] David I Holmes, Michael Robertson, and Roxanna Paez. Stephen crane and the new-york tribune: A case study in traditional and non-traditional authorship attribution. *Computers and the Humanities*, 35(3):315–331, 2001.

[42] David I Holmes and Fiona Tweedie. Forensic stylometry: A review of the cusum controversy. *Revue Informatique et Statistique dans les Sciences Humaines*, 31:19–47, 1995.

[43] Johan F Hoorn, Stefan L Frank, Wojtek Kowalczyk, and Floor van Der Ham. Neural network identification of poets using letter sequences. *Literary and Linguistic Computing*, 14(3):311–338, 1999.

[44] David L Hoover. Frequent word sequences and statistical stylistics. *Literary and Linguistic Computing*, 17(2):157–180, 2002.

[45] David L Hoover. Another perspective on vocabulary richness. *Computers and the Humanities*, 37(2):151–178, 2003.

[46] David L Hoover. Frequent collocations and authorial style. *Literary and Linguistic Computing*, 18(3):261–286, 2003.

[47] David L Hoover. Multivariate analysis and the study of style variation. *Literary and Linguistic Computing*, 18(4):341–360, 2003.

[48] David L Hoover. Delta prime? *Literary and Linguistic Computing*, 19(4):477–495, 2004.

[49] David L Hoover. Testing burrows's delta. *Literary and Linguistic Computing*, 19(4):453–475, 2004.

[50] Patrick Juola. Authorship attribution. *Foundations and Trends in information Retrieval*, 1(3):233–334, 2006.

[51] Patrick Juola and Harald Baayen. A controlled-corpus experiment in authorship identification by cross-entropy. *Literary and Linguistic Computing*, 20(Suppl):59–67, 2005.

[52] Ryan Kelly. pyenchant: a spellchecking library for python, 2014. [Online; accessed 2014-11-12].

[53] Vlado Kešelj, Fuchun Peng, Nick Cercone, and Calvin Thomas. N-gram-based author profiles for authorship attribution. In *Conference Pacific Association for Computational Linguistics*, volume 3, pages 255–264, 2003.

[54] Dmitri V Khmelev and Fiona Tweedie. Using markov chains for identification of writer. *Literary and Linguistic Computing*, 16(4):299–307, 2002.

[55] Dmitry V Khmelev and William J Teahan. A repetition based measure for verification of text collections and for text categorization. In *International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 104–110. ACM, 2003.

[56] Bradley Kjell. Authorship attribution of text samples using neural networks and bayesian classifiers. In *IEEE International Conference on Systems, Man, and Cybernetics, 1994. Humans, Information and Technology.*, volume 2, pages 1660–1664. IEEE, 1994.

[57] Bradley Kjell. Authorship determination using letter pair frequency features with neural network classifiers. *Literary and Linguistic Computing*, 9(2):119–124, 1994.

[58] Bradley Kjell, Willis A Woods, and Ophir Frieder. Discrimination of authorship using visualization. *Information processing & management*, 30(1):141–150, 1994.

[59] Bradley Kjell, Willis A Woods, and Ophir Frieder. Information retrieval using letter tuples with neural network and nearest neighbor classifiers. In *IEEE International Conference on Systems, Man and Cybernetics, 1995. Intelligent Systems for the 21st Century.*, volume 2, pages 1222–1226. IEEE, 1995.

[60] Moshe Koppel and Jonathan Schler. Exploiting stylistic idiosyncrasies for authorship attribution. In *Workshop on Computational Approaches to Style Analysis and Synthesis*, volume 69, pages 72–80, 2003.

[61] Moshe Koppel, Jonathan Schler, and Shlomo Argamon. Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, 60(1):9–26, 2009.

[62] Moshe Koppel, Jonathan Schler, Shlomo Argamon, and Eran Messeri. Authorship attribution with thousands of candidate authors. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 659–660. ACM, 2006.

[63] Moshe Koppel, Jonathan Schler, and Kfir Zigdon. Determining an author's native language by mining a text for errors. In *ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pages 624–628. ACM, 2005.

[64] Raffi Krikorian. New tweets per second record, and how! Twitter Blog, 2013. Accessed on May, 2014 on `http://tinyurl.com/kcuhdcw`.

[65] OV Kukushkina, AA Polikarpov, and Dmitry Viktorovich Khmelev. Using literal and grammatical statistics for authorship attribution. *Problems of Information Transmission*, 37(2):172–184, 2001.

[66] Robert Layton, Stephen McCombie, and Paul Watters. Authorship attribution of irc messages using inverse author frequency. In *Cybercrime and Trustworthy Computing Workshop (CTC), 2012 Third*, pages 7–13. IEEE, 2012.

[67] Robert Layton, Paul Watters, and Richard Dazeley. Authorship attribution for twitter in 140 characters or less. In *Cybercrime and Trustworthy Computing*, pages 1–8, 2010.

[68] Gerard Ledger and Thomas Merriam. Shakespeare, fletcher, and the two noble kinsmen. *Literary and Linguistic Computing*, 9(3):235–248, 1994.

[69] Jiexun Li, Rong Zheng, and Hsinchun Chen. From fingerprint to writeprint. *Communications of the ACM*, 49(4):76–82, 2006.

[70] David Lowe and Robert Matthews. Shakespeare vs. fletcher: A stylometric analysis by radial basis functions. *Computers and the Humanities*, 29(6):449–461, 1995.

[71] David Madigan, Alexander Genkin, David D Lewis, Shlomo Argamon, Dmitriy Fradkin, and Li Ye. Author identification on the large scale. In *Meeting of the Classification Society of North America*, 2006.

[72] Christopher D Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[73] David Mannion and Peter Dixon. Authorship attribution: the case of oliver goldsmith. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46(1):1–18, 1997.

[74] Colin Martindale and Dean McKenzie. On the utility of content analysis in author attribution: The federalist. *Computers and the Humanities*, 29(4):259–270, 1995.

[75] Conrad Mascol. Curves of pauline and pseudo-pauline style i. *Unitarian Review*, 30:453–460, 1888.

[76] Robert AJ Matthews and Thomas VN Merriam. Neural computation in stylometry i: An application to the works of shakespeare and fletcher. *Literary and Linguistic Computing*, 8(4):203–209, 1994.

[77] David Mealand. Measuring genre differences in mark with correspondence analysis. *Literary and Linguistic Computing*, 12(4):227–245, 1997.

[78] David L Mealand. Correspondence analysis of luke. *Literary and Linguistic Computing*, 10(3):171–182, 1995.

[79] Thomas C Mendenhall. The characteristic curves of composition. *Science*, pages 237–246, 1887.

[80] Thomas Merriam. What shakespeare wrote in henry viii (part i). *The Bard*, 2(1979):81–94, 1979.

[81] Thomas Merriam. What shakespeare wrote in henry viii (part ii). *The Bard*, 2(1980):11–n8, 1980.

[82] Thomas Merriam. The authorship of sir thomas more. *ALLC Bulletin. Association for Library and Linguistic Computing Bangor*, 10(1):1–7, 1982.

[83] Thomas Merriam. An investigation of morton's method: A reply. *Computers and the Humanities*, 21:57–8, 1987.

[84] Thomas Merriam. Marlowe's hand in edward iii revisited. *Literary and Linguistic Computing*, 11(1):19–22, 1996.

[85] Thomas V N Merriam and Robert A J Matthews. Neural computation in stylometry II: An application to the works of Shakespeare and Marlowe. *Literary and Linguistic Computing*, 9(1):1–6, 1994.

[86] George K Mikros and Kostas Perifanos. Authorship attribution in greek tweets using authors multilevel n-gram profiles. In *AAAI Spring Symposium Series*, 2013.

[87] Robert Moore. Twitter data analysis: An investor's perspective. Techcrunch blog, October 2009. Accessed on May, 2014 on `http://techcrunch.com/2009/10/05/twitter-data-analysis-an-investors-perspective-2/`.

[88] Andrew Q Morton. The authorship of greek prose. *Journal of the Royal Statistical Society. Series A (General)*, pages 169–233, 1965.

[89] Andrew Q Morton. *Literary detection: How to prove authorship and fraud in literature and documents*. Bowker London, 1978.

[90] Andrew Q Morton. Once. a test of authorship based on words which are not repeated in the sample. *Literary and Linguistic Computing*, 1(1):1–8, 1986.

[91] Andrew Q Morton. *Proper Words in Proper Places: A General Introduction to the Use of Cumulative Sum Techniques for Identifying the Source of Written Or Spoken Utterance*. Department of Computing Science, University of Glasgow, 1991.

[92] Andrew Q Morton and Sidney Michaelson. *The qsum plot*, volume 3. University of Edinburgh, Department of Computer Science, 1990.

[93] Frederick Mosteller and David L Wallace. *Inference and Disputed Authorship: The Federalist Papers*. Addison-Wesley, Reading, Mass., 1964.

[94] Daniel Pavelec, Edson JR Justino, and Luiz S Oliveira. Author identification using stylometric features. *Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial*, 11(36):59–66, 2007.

[95] Fuchun Peng, Dale Schuurmans, and Shaojun Wang. Augmenting naïve bayes classifiers with statistical language models. *Information Retrieval*, 7(3-4):317–345, 2004.

[96] Fuchun Peng, Dale Schuurmans, and Shaojun Wang. Augmenting naïve bayes classifiers with statistical language models. *Information Retrieval*, 7(3-4):317–345, 2004.

[97] Phi-Long. guess_language: Guess the natural language of a text, 2014. [Online; accessed 2014-11-12].

[98] Emily Ramshaw. Bashing the candidates with their own names. The New York Times, May 2012. Accessed on May, 2014 on `http://tinyurl.com/q6lc2fw`.

[99] Michael G Reed, Paul F Syverson, and David M Goldschlag. Anonymous connections and onion routing. *IEEE Journal on Selected Areas in Communications*, 16(4):482–494, 1998.

[100] Gerard Salton and Michael J McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986.

[101] Conrad Sanderson and Simon Guenter. Short text authorship attribution via sequence kernels, markov chains and author unmasking: An investigation. In *Conference on Empirical Methods in Natural Language Processing*, pages 482–491. Association for Computational Linguistics, 2006.

[102] Anthony J Sandford, Joy P Aked, Linda M Moxey, and James Mullin. A critical examination of assumptions underlying the cusum technique of forensic linguistics. *Forensic Linguistics*, 1(2):151–167, 1994.

[103] Roy Schwartz, Oren Tsur, Ari Rappoport, and Moshe Koppel. Authorship attribution of micro-messages. In *Conference on Empirical Methods on Natural Language Processing*, pages 1880–1891. ACL, 2013.

[104] Gerry Shih. Anonymous twitter feeds arise as political weapon. The New York Times, June 2014. Accessed on May, 2014 on `http://tinyurl.com/5vol3gt`.

[105] Rui S Silva, Gustavo Laboreiro, Luís Sarmento, Tim Grant, Eugénio Oliveira, and Belinda Maia. 'twazn me!!!;('automatic authorship analysis of micro-blogging messages. In *Natural Language Processing and Information Systems*, pages 161–168. Springer, 2011.

[106] Michael W A Smith. An investigation of morton's method to distinguish elizabethan playwrights. *Computers and the Humanities*, 19(1):3–21, 1985.

[107] Michael W A Smith. An investigation of the basis of morton's method for the determination of authorship in readers and authors. *Style*, 19(3):341–368, 1985.

[108] Michael W A Smith. Merriam's applications of morton's method. *Computers and the Humanities*, 21(1):59–60, 1987.

[109] Harold Somers. An attempt to use weighted cusums to identify sublanguages. In *Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning*, pages 131–139. Association for Computational Linguistics, 1998.

[110] Harold Somers and Fiona Tweedie. Authorship attribution and pastiche. *Computers and the Humanities*, 37(4):407–429, 2003.

[111] Efstathios Stamatatos. Author identification: Using text sampling to handle the class imbalance problem. *Information Processing & Management*, 44(2):790–799, 2008.

[112] Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, 2009.

[113] Efstathios Stamatatos. On the robustness of authorship attribution based on character n-gram features. *Journal of Law & Policy*, 21:421–725, 2013.

[114] Efstathios Stamatatos, Nikos Fakotakis, and George Kokkinakis. Automatic text categorization in terms of genre and author. *Computational linguistics*, 26(4):471–495, 2000.

[115] Efstathios Stamatatos, Nikos Fakotakis, and Georgios Kokkinakis. Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35(2):193–214, 2001.

[116] Emily K Tse, Fiona Tweedie, and Bernard D Frischer. Unravelling the purple thread: Function word variability and the scriptores historiae augustae. *Literary and Linguistic Computing*, 13(3):141–149, 1998.

[117] Fiona Tweedie, David I Holmes, and Thomas N Corns. The provenance of de doctrina christiana, attributed to john milton: a statistical investigation. *Literary and Linguistic Computing*, 13(2):77–87, 1998.

[118] Fiona Tweedie, Sameer Singh, and David I Holmes. Neural network applications in stylometry: The federalist papers. *Computers and the Humanities*, 30(1):1–10, 1996.

[119] Hans Van Halteren. Linguistic profiling for author recognition and verification. In *Annual Meeting on Association for Computational Linguistics*, page 199. Association for Computational Linguistics, 2004.

[120] William C Wake. Sentence-length distributions of greek authors. *Journal of the Royal Statistical Society. Series A (General)*, pages 331–346, 1957.

[121] Jennifer Waters. Why id thieves love social media. The Wall Street Journal, March 2012. Accessed on May, 2014 on `http://tinyurl.com/ldvhpsb`.

[122] Sam Waugh, Anthony Adams, and Fiona Tweedie. Computational stylistics using artificial neural networks. *Literary and Linguistic Computing*, 15(2):187–198, 2000.

[123] Jianxin Wu. Power mean svm for large scale visual classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2344–2351, 2012.

[124] George U Yule. On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship. *Biometrika*, pages 363–390, 1938.

[125] George Udny Yule. *The Statistical Study of Literary Vocabulary*. Cambridge University Press Archive, 1944.

[126] Ying Zhao and Justin Zobel. Effective and scalable authorship attribution using function words. In *Asian Information Retrieval Symposium*, pages 174–189. Springer, 2005.

[127] Ying Zhao and Justin Zobel. Searching with style: Authorship attribution in classic literature. In *Australasian Conference on Computer Science*, volume 62, pages 59–68. Australian Computer Society, Inc., 2007.

[128] Ying Zhao, Justin Zobel, and Phil Vines. Using relative entropy for authorship attribution. In *Information Retrieval Technology*, pages 92–105. Springer, 2006.

[129] Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan Huang. 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, 2006.

[130] George K Zipf. *Selected studies of the principle of relative frequency in language*. Harvard Univ. Press, 1932.

Appendix A

List of the function words used

- a
- about
- above
- after
- all
- although
- am
- among
- an
- and
- another
- any
- anybody
- anyone
- anything
- are
- around
- as
- at
- be
- because
- before
- behind
- below
- beside
- between
- both
- but
- by
- can
- cos
- do
- down
- each
- either
- enough
- every
- everybody
- everyone
- everything
- few
- following
- for
- from
- have
- he
- her
- him
- i
- if
- in
- including
- inside
- into
- is
- it
- its
- latter
- less
- like
- little
- lots
- many
- me
- more
- most
- much
- must
- my
- near
- need
- neither
- no
- nobody
- none
- nor
- nothing
- of
- off
- on
- once
- one
- onto
- opposite

- or
- our
- outside
- over
- own
- past
- per
- plenty
- plus
- regarding
- same
- several
- she
- should
- since
- so
- some
- somebody
- someone
- something
- such
- than
- that
- the
- their
- them
- these
- they
- this
- those
- though
- through
- till
- to
- toward
- towards
- under
- unless
- unlike
- until
- up
- upon
- us
- used
- via
- we
- what
- whatever
- when
- where
- whether
- which
- while
- who
- whoever
- whom
- whose
- will
- with
- within
- without
- worth
- would
- yes
- you