Faculty of Computer Science and Information Technology

Author Identification of English Tweets for Social Media Forensics

Nursyahirah Binti Tarmizi

Master of Science
2023

Author Identification of English Tweets for Social Media Forensics

Nursyahirah Binti Tarmizi

A thesis submitted

In fulfillment of the requirements for the degree of Master of Science

(Language Technologies)

Faculty of Computer Science and Information Technology
UNIVERSITI MALAYSIA SARAWAK
2023

DECLARATION

I declare that the work in this thesis was carried out in accordance with the regulations of Universiti Malaysia Sarawak. Except where due acknowledgements have been made, the work is that of the author alone. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Syira

………………………………

Signature

Name:	Nursyahirah Binti Tarmizi
Matric No.:	17020134

Faculty of Computer Science and Information Technology

Universiti Malaysia Sarawak

Date: 24th August 2023

ACKNOWLEDGEMENT

In the name of Allah the most Gracious, the most Merciful, all praises to Allah S.W.T for this wonderful journey. I’m grateful for the utmost support from my family, especially my parents Harimah and Tarmizi as well as my sister Syazana, my aunts, uncles especially Masni and Hambaly, and cousins for their continuous love and du’a during the creation of my thesis. Without you I would not be where I am today.

I am very thankful to my supervisor, Dr Suhaila Saee and co-supervisor, Dr Dayang Hanani Abang Ibrahim, who supervised me throughout my studies and always support me in my own research ideas. Even in the busiest times, feedbacks and supports were always available and the discussion of ideas helped me a lot.

I would also like to thank all my colleagues at FCSIT especially Postgraduate Lab 1, whom I have honour to change ideas and opinions on a regular basis. Besides discussing ideas and changing opinions, colleagues like them cannot be taken for granted and are among the reasons why I enjoyed this period in my life as much as I did.

My sincere gratitude to the Centre for Graduate Studies, for the advice and support given during my period of study in Universiti Malaysia Sarawak.

Finally, I would like to thank the management of the Universiti Malaysia Sarawak for making it possible for me to complete my study here in Sarawak. Thank you all.

ABSTRACT

Authorship Identification (AI) is the process of determining the most likely author of a given text by analysing writing style characteristics and linguistic patterns. Identifying the author of online social network (OSN) text becomes a pressing issue nowadays as the increase of cyberbully cases among the social media users. AI plays vital role in social media forensics (SMF) to unveil the true identity of the cyberbullying perpetrator from the OSN text. However, OSN text has been an open problem in AI as the limited length of the text and the usage of Internet jargon affecting the performance of AI system. In this research, AI task is conducted to facilitate the SMF activity by analysing the writing style of tweets from Twitter in identifying most plausible author for anonymized tweet. The writing style of the author or known as the stylometric features including character n-grams, word n-grams and Part-of-Speech (POS) n-grams are extracted from the text. These features are used widely in identifying the author of short text as they are language independent and tolerant of grammatical errors. The features are represented using different text representation models namely TF-IDF and Embedding model. The models are examined to compare which one could best represent the OSN text. For classification, machine learning and deep learning are used to evaluate the classification model by maintaining the optimum performance of AI system. The findings shown that Twitter native features are very useful in boosting the performance of AI system. Embedding-based model achieved better performance in representing n-grams with fix and distributed representation. The best result was achieved when CNN mix with embedding-based model with accuracy of 95.02% for English and 94% for KadazanDusun and both 95 % precision for both languages.

Keywords: Author Identification, authorship analysis, cyberbullying, stylometry, tweets

Identifikasi Penulis Tweets Bahasa Inggeris untuk Forensik Media Sosial

ABSTRAK

Identifikasi penulis adalah proses untuk mengenalpasti penulis yang paling munasabah bagi sesuatu text dengan menganalisa ciri gaya penulisan dan corak linguistik. Pengenalpastian identiti penulis text media sosial menjadi isu yang serius pada masa kini akibat peningkatan kes jenayah pembuli siber dalam kalangan pengguna media sosial. Identifikasi penulisan memainkan peranan yang penting dalam forensik media sosial untuk mendedah pemilik sebenar teks pembulian siber. Walau bagaimanapun, teks media sosial yang pendek merupakan masalah semasa dalam identifikasi penulisan kerana ukuran teks yang terhad dan penggunaan bahasa Internet yang tidak teratur akan menjatuhkan prestasi sistem pengenalpastian identiti penulis. Dalam kajian ini, tugasan AI dijalankan bagi membantu aktiviti SMF dengan menganalisa bentuk penulisan tweets daripada Twitter bagi mengenalpasti identiti penulis tweet tanpa nama. Gaya penulisan atau dikenali sebgai ciri-ciri stylometrik, seperti n-gram abjad, n-gram perkataan dan n-gram Part-of- Speech (POS) akan diekstrak. Ciri-ciri stylometrik ini digunakan secara meluas dalam mengenalpasti penulis teks media sosial kerana ciri-cirinya yang bersifat berdikari bahasa dan dapat bertolak ansur dengan kesalahan tatabahasa. Dua model perwakilan teks, TF-IDF dan Embedding telah digunakan dan dibandingkan keduanya untuk meneliti yang mana antara dua model tersebut lebih baik dalam mewakili teks media sosial. Untuk proses klasifikasi, pendekatan machine learning dan deep learning telah digunakan untuk menilai algoritma klassifikasi yang mana satu dapat mengekalkan prestasi optimum bagi sistem pengenalpastian penulis text media sosial. Dapatan kajian menunjukkan bahawa ciri-ciri asli Twitter sangat berguna dalam meningkatkan prestasi system AI. Dapatan juga menunjukkan model berasaskan embedding lebih baik dalam mewakili ciri-ciri n-grams *dengan perwakilan tetap dan teragih. Keputusan terbaik dicapai apabila model CNN bergabung dengn gabungan embedding dengan ketepatan 95.02% untuk Bahasa Inggeris dan 94% untuk KadazanDusun dan keduanya mencapai 95% kepersisan.*

Kata kunci: *Analisa identifikasi penulisan, pembulian siber, identifikasi penulisan, stilometri, tweets*

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

AI	Author Identification
AV	Author Verification
AP	Author Profiling
CNN	Convolutional Neural Network
DFD	Digital Forensics Department
DNN	Deep Neural Network
DL	Deep Learning
ML	Machine Learning
OSN	Online Social Network
TF-IDF	Term Frequency – Index Document Frequency
TLS	Transport Layer Security
TOR	The Onion Router
SMF	Social Media Forensics
SVM	Support Vector Machines
Word2Vec	Word-to-Vector
U-RL	Under-Resourced Language

CHAPTER 1

INTRODUCTION

Chapter 1 provides an overview of authorship analysis, including the major tasks, data types, and processes involved. It also discusses the research motivation, questions, problems, objectives, methodology, significance, and contributions of the study. The chapter concludes with a summary of the thesis structure. In the rest of the thesis, the term ‘AI’ is referred as Authorship Identification task.

1.1 Authorship Analysis

Authorship analysis is a process of examining the characteristics of a piece of writing to draw conclusions on its authorship. Authorship analysis has been used in a small but diverse number of application areas including identifying authors in literature, in the program code, and in forensic analysis for criminal cases (De Vel et al., 2001). Authorship analysis has rooted in the authorship attribution problem of historical literature such as, resolving the debate on Shakespeare’s work, solving the author debates over the *Federalist Papers* and *Unabomber Manifesto* (Zheng et al., 2003). Figure 1.1 illustrates the general pipeline of authorship analysis which describes the major authorship tasks, scenarios involved, types of data used, and the general process involved.

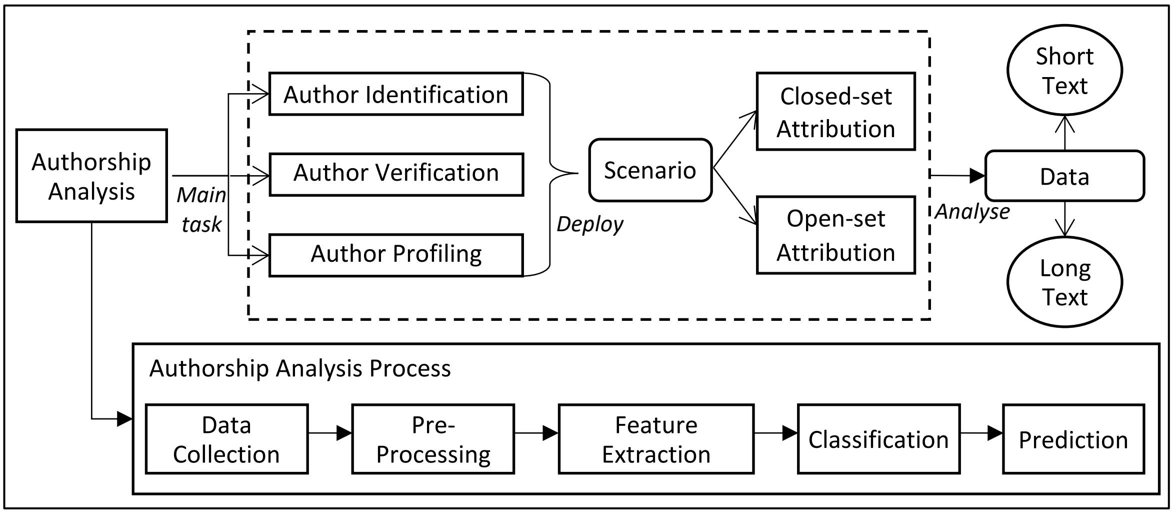

Figure 1.1: A general pipeline of authorship analysis

Referring to Figure 1.1, there are three main tasks in authorship analysis. The tasks consist of processes in determining the authorship information of the anonymous text. The three tasks namely Author Identification (AI), Author Verification (AV) and Author Profiling (AP), may be deployed in one of the two scenarios. The scenarios are called the closed-set and open-set attribution. There are two types of data that are used in gathering the internal evidence of the writing style. The general process of authorship analysis fits the standard modern paradigm of text classification problem (Koppel et al., 2009) where the components of text classification are used in authorship analysis process.

Authorship analysis involves evaluating the writing characteristics to make inferences about who wrote it. Kešelj et al. (2003) states that an author’s writing style is formed by extracting stylometric features from their own texts. According to De Vel et al. (2001) stylometric features or "style markers" are used in early authorship attribution studies to define the features and quantifying the writing style of an author. In this report, the term stylometric features will be used in the explanations.

As an example, the style of an author can be represented by the choice and use of words, organisation of the sentences and paragraph, and the use of function and content words. Prior to the stylometric feature, a line of research known as ‘stylometry’ attempts to define the features for quantifying the writing style (Stamatatos, 2008). Stylometry is a linguistic discipline that applies statistical analysis to literary style (Diurdeva et al., 2016).

1.1.1 Major Tasks in Authorship Analysis

Authorship analysis task tackles the problem of determining the author of anonymous text or text that the authorship is open to dispute. Through authorship analysis task, one can solve the mystery of suicide note, ransom note or even threatening email by identifying the real culprit and release the innocent people which work close to forensic investigation and criminal law. As for the domain humanities, one would like to solve the attribution of the disputed document or reveal the anonymous author that utilising the historical documents or literary documents. There are three major tasks in authorship analysis as shown in Figure 1.2 followed by the explanation of each task.

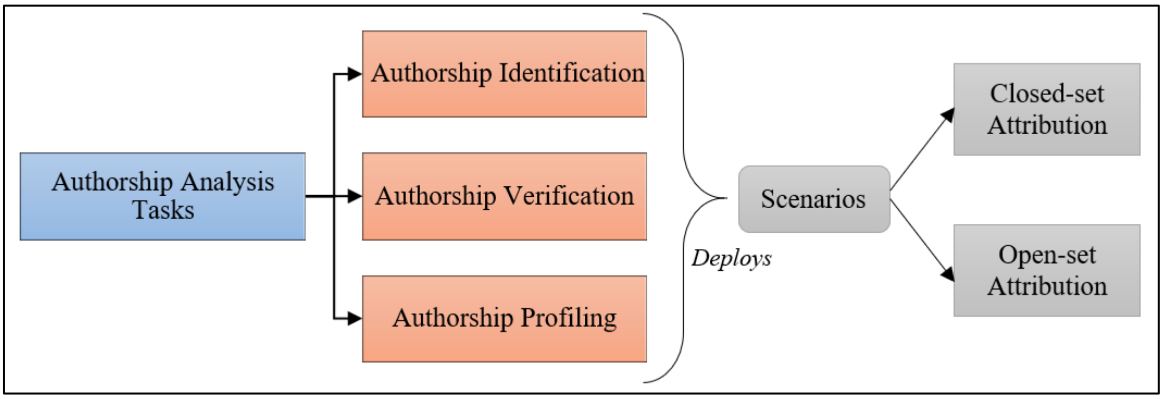

Figure 1.2: Major tasks in Authorship Analysis

Based on Figure 1.2, Authorship Identification (AI) aims to identify the author of a given document from a set of suspects by examining the other writings by that author. De Vel et al. (2001) states that the goal of AI task is to predict the author of the text selected from a pool of potential candidates. The task is called as authorship (or author) identification or authorship attribution by researchers with a background in computer science (Stamatatos, 2009).

Author Verification (AV) aims to determine whether the examined text is also written by the same author as reported by Stamatatos (2017). Rather than trying to select the most plausible author among a given number of candidates, the classification process of AV task consists of two possible outcomes either yes (indeed the person wrote the examined text) or no (the text was not written by the person).

Author Profiling (AP) is done by extracting information about the age, education, gender, etc. of the author of a given text (Koppel et al., 2002). Author profiling is only limited to determine the author's demographic including gender, age class, and native language (Soler-Company & Wanner, 2017). Besides, the profile of an author can also be the author's psychological traits (Argamon et al., 2009).

According to Figure 1.2, the authorship analysis tasks can be carried out based on two types of attribution scenarios. In the close-set scenario, a text of unknown authorship is attributed to one candidate author, given a well-defined list of candidate authors with the sample of texts they authored (Potthast et al., 2019). While, in the open-set scenario, several varieties of attribution problems fall under this scenario such as:

- There is no candidate set at all. In this case, the challenge is to provide as much as demographic or psychological information about the author.

- Needle-in-a-haystack problem happened when there are thousands of candidate authors but with limited text samples provided.

- The true author may not be included in the list of suspects and the challenge is to determine if the author is the owner of the text or otherwise.

1.1.2 Types of Data in Authorship Analysis

Figure 1.3 below depicts different types of data that are used as the evidence to be analysed in any authorship task.

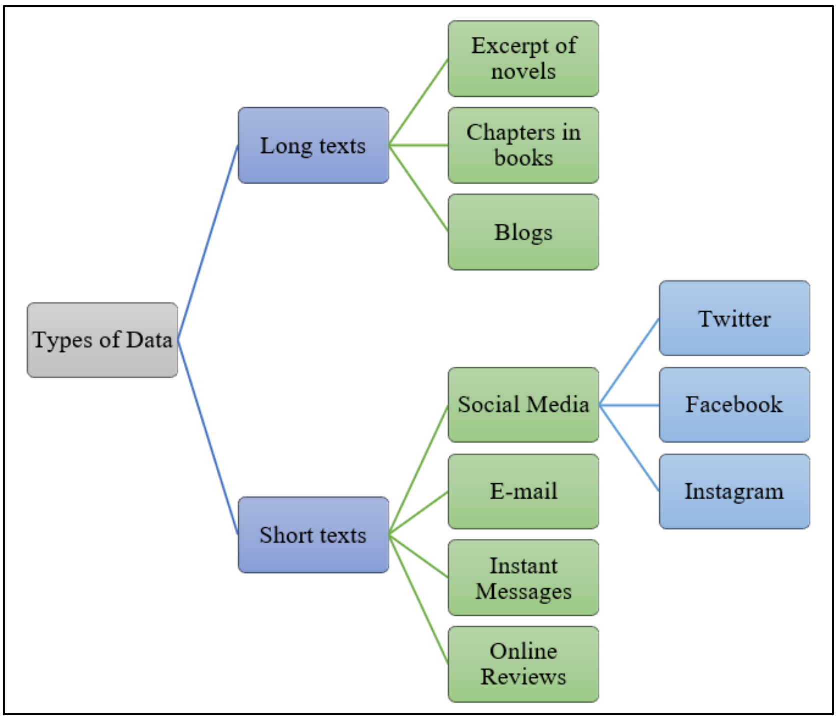

Figure 1.3: Types of data in Authorship Analysis

Based on Figure 1.3, there are two types of data that are being analysed in authorship analysis, which are short text and long text. Long texts data composed of texts from novel excerpts and books for historical literature study (Gladwin et al., 2017). However, the extensive amount of electronic texts available through Internet media (blogs and web forum) has increased the need for handling the text efficiently. Long text has been used to analyse the writing style of terrorist groups (Abbasi & Chen, 2005) for forensic study to track the communication of militant groups and terrorist organizations using e-mails as evidence.

Presently, with the evolution of the Internet and information technology, people often use online communication to interact and socialize with each other through networks. Digital documents that are produced through the Internet are handy and essential to perform authorship analysis on short texts for the purpose of forensic investigation (Banga & Mehndiratta, 2017; Okuno et al., 2015; Orebaugh & Allnutt, 2009). Short texts such as Instant Messaging (IM), social media (Facebook, Twitter, Instagram etc.) and reviews of online products, produce massive digital documents.

Unfortunately, the evolution and divergence of the Internet are often being used negatively on online activities. Cyber-crimes such as Internet scams, stealing identity and cyberbullying (twitter, 2013) keep growing each day. According to (Diurdeva et al., 2016), the usage of digital documents processing method is crucial nowadays to analyse short text document in solving cybercrime issues.

However, AI for short text has always been difficult compared to long text. Practically, the task of extracting stylometry features easier for long texts as there are many textual features can be extracted for each author. Whereas short texts are harder to deal because of the text-length limitation and insufficient content. Thus, the limited amount of text-length and content make the identification process much harder and complicated for short text.

1.1.3 General process of Authorship Analysis

Authorship analysis has adopted the general process of text classification. In the simplest form of problem, authorship analysis task is to predict the author of disputed text given the examples of the writing of candidate authors. This makes authorship analysis task fits the standard modern paradigm of a text classification process. Authorship analysis studies differ in terms of the stylometric features used and the technique used to extract the features and classify the author's text. Figure 1.4 below shows the general process of authorship analysis.

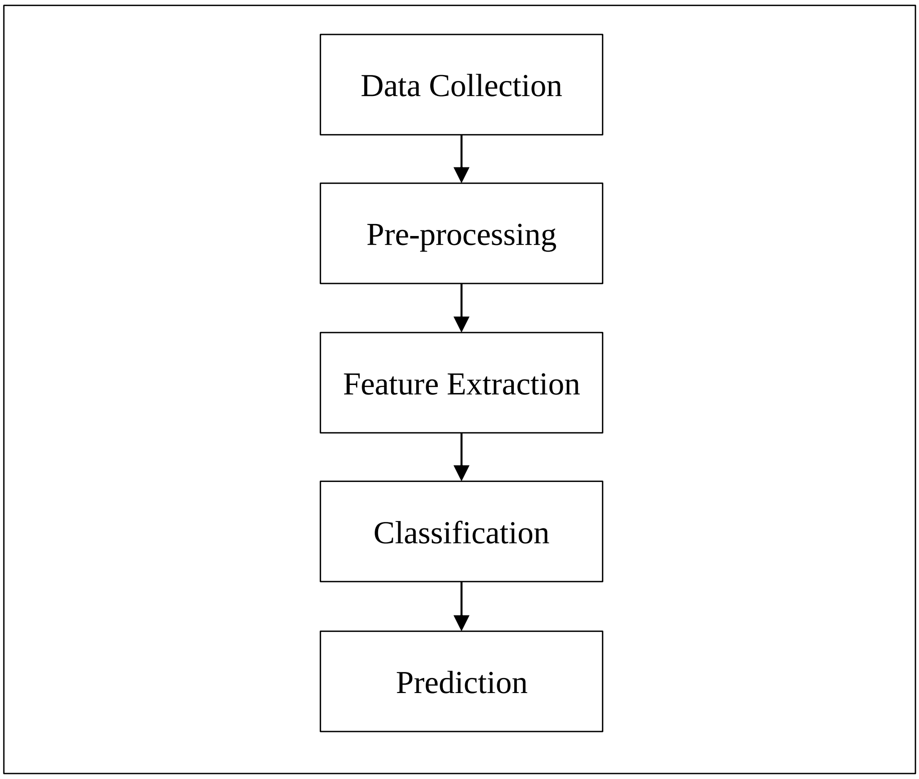

Figure 1.4: The general process of Authorship Analysis

Based on Figure 1.4, the first step of authorship analysis is data collection. Data is the text collected for each author. After collecting a sufficient amount of text for each author, the next step is pre-processing. In pre-processing, the raw version of the text is cleaned from meta-data and analysed by using a text analysis tool to extract the textual information. Then, the textual information is extracted as stylometric features and converted into vectors. Later, the features are learned by the classification algorithms (e.g., Naïve Bayes, Neural Networks, and Decision Trees). Lastly, the learned models will be used to predict the author of anonymous text.

As stated by (Markov, 2017), from a machine learning perspective, approaches to AI task can be viewed as a multi-class single-label text classification problem. The problem includes a set of class labels which are known as prior. While, for AV task, the classification of data involves binary classification which verify the documents do belong to the author or otherwise. Similar to AV, AP consists of determining the demographic and psychological characteristics of the author of the anonymous document. For instance, to determine whether the author is male or female.

1.2 Research Motivation

Social media is increasingly popular and has become a good socialising tool where people can get connected without any limitations as of distance and times. Social media facilitates the creation and sharing of various form of content such as blogs, videos and photos. However, the usage of online activities via social media often being used negatively. Cyber-criminal activities such as social media cyberbullying keep growing each day (Bhargava et al., 2013).

Moreover, recent technology called TOR (The Onion Routing) network provides a service that encrypts user data and randomly sends back and forth to various nodes of anonymisation. For instance, one can simply tunnel traffics through a series of proxy server, thus rendering a harder way to locate the originating IP address. The advanced network technology may create frustration on network forensics to track down the criminal where the online identity and the location of the criminal are invisible. In such case, the text left on social media may be the only clue to the author's identity.

Nowadays, there are organizations handle cyber-criminalistics investigations who reveal the identity of the anonymous user on the internet for criminal prosecution purposes. Figure 1.5 below depicts the authorship identification task in the digital forensics' investigation case.

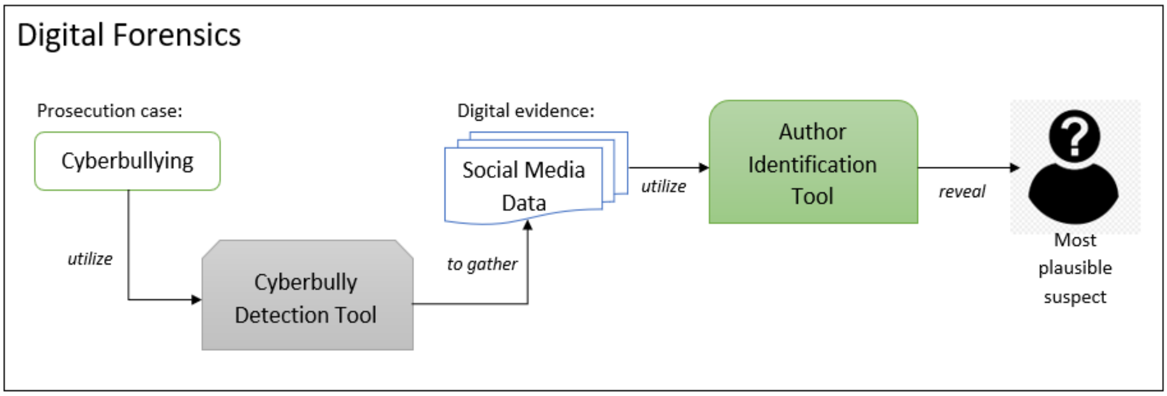

Figure 1.5: The utilization of AI tool in digital foreniscs case

In Figure 1.5, AI task is being carried out in digital forensics to facilitate further the prosecution by identifying the most plausible suspect that conducted the cyberbully activity in social media. Henceforth, the importance of authorship analysis in digital forensic investigation necessitates the needs for author identification task. The task of AI in social media forensics is decisive in applications of cybersecurity, social media analytic and digital humanities.