A Study on Author Identification through Stylometry

Lakshmi
M.Tech Student (Computer Science)
Lovely Professional University
Phagwara, India
erlakshmi.gosain@gmail.com

Pushpendra Kumar Pateriya
Assistant Professor (Computer Science)
Lovely Professional University
Phagwara, India
pushpendra.mnnit@gmail.com

Abstract

Electronic communication is one of the popular ways of communication in this era. E-mail communication is the most popular way of electronic communication. Internet works as the backbone for these communications. In digital forensics, questions is arises that the authors of documents and the author identity, demographic background is linked to other documents or not. So identification of the author(s) of the message(s) and non repudiation are some of the major challenges. Author identification is a critical point to be ensured, because many people are used to copy the content of others. Stylometry can be used for the author identification for text documents. As the non–repudiation and integrity of the message are the major concerns. Stylometry is not only identifying a writing pattern but we can also identify the gender of the human. So this document discussed about identification of author, authentication through stylometry technique. In this paper different stylometric techniques are discussed.

***Keywords:* Stylometry, author identification, email, gender.**

1. Introduction

In 1851 the use of tools i.e. statistical tools to test questions of authorship was done when mathematician Augustus de Morgan proposed using average word length to numerically characterize authorship style [1]. After that Thomas Mendenhall who was a physicist and proposed that an author has a "characteristic curve of composition" determined by how an author uses words of different lengths frequently, in year 1887 [2]. In 1888 a mathematician (William Benjamin Smith) published two papers describing a "curve of style" to distinguish authorial styles based on average sentence lengths, this technique was applied to the Pauline Epistles [3]. A book "Principes de stylometrie" 1890 was given by the Polish philosopher Wincenty Lutosławski to describe the basics of stylometry. Chronology of Plato's Dialogues was given by Lutosławski by using this method. Then Lucius Sherman, a professor of English in 1893, found that writing style over time changes with average sentence length [4]. Due to the increasing computing power, availability of the Internet, growth of ultrahigh dimensional statistical tools the stylometric techniques are growing rapidly day by day. In this paper, basically we focused on the various types of stylometry techniques. This paper is organized as follows: in section 2; we have the description of stylometry and in section 3; we have discussed about the literature Review. In the section 4, is providing the comparative analysis of research based on scientific articles, e-mail author identification using stylometry and in the end in section 4, is providing the conclusion over the discussion given in the paper.

2. Stylometry

Stylometry is a kind of study by which a person can judge about another person by its writing style. An example is discussed; any experienced can apply a kind of stylometry. Which of the example was written more recently? Are there are two authors or only one? Which example was written by a native English speaker?

For example:-

Table 1: Stylometry example

Example A	*Example B*
I am not able to give you money. I am going to home. Are you coming there?	I will be trying to provide all kinds of notes that can be related to study or the internet.

Stylometry is used for detect the plagiarism, which is a very serious issue in education system. Stefan Gruber and Stuart Noven proposed a software tool that support detection of plagiarism in 2005 [5]. For Communication purpose now a day, the messages are passed through e-mail. The misuse of e-mail is increasing day by day. The people are doing crime by emails and they send spam messages, hoaxes and threats. Therefore, it is important to properly identify the author of the e-mail. There has been growing interest in applying stylometry to the content generation where the content is checked whether it is original or copied from others style. An example of an e-mail hoax is sending a false computer virus warning with the request to send the warning on to all the recipients, thus the mail server time and bandwidth would be wasted. Computer viruses or worms are now commonly distributed by e-mail, by making use of loose security features in some e-mail programs. These worms copy themselves to all of the addresses in the recipient's address book. Author identification have grown in several different areas in past years in practical manner such as, civil law in which identification of copyright and estate disputes, criminal law in which identification of writers of ransom notes and harassing letters, and computer security in which mining email content are identified [6]. The analysis of the texts for evidence of authenticity, authorial identity has also increased the stylometry techniques. The English professor John Burrows concluded that the intellectual propensities of the author's display inherently and written texts have a particular style. If we don't know the authorship of the word-use patterns in a text and then comparing and contrasting those patterns to the patterns in texts of known authorship, the similarity and dissimilarities of the textual patterns can provide supporting evidence for or contradicting evidence against an assertion of authorship [7].

3. Literature Review

The fields of stylistics, computational linguistics, and non-traditional authorship attribution to develop a possible framework for the identification of e-mail text authorship. The fields like text classification, machine learning, software forensics, and forensic linguistics also impact on the current study. Plagiarism detection [8] can be seen as complementary to stylometric authorship attribution: it attempts to detect common content between documents, even if the style may have been changed. Authorship attribution and authorship characterization are quite distinct problems from plagiarism detection. Authorship analysis has been used in a number of application areas such as identifying authors in literature, in program code, etc. In the authorship attribution literature there are three kinds of evidence that can be used to establish authorship i.e. external, interpretive and linguistic.

- External evidence includes the handwriting or a signed manuscript of author.
- Interpretive evidence is the study of document i.e. when it was written, what the author meant by it and how that can be compared to other works by the same author.
- Linguistic evidence is focusing on the patterns of words and the actual words that are used in a document.

In some domains, statistical techniques have successfully deduced author identity. Stylometric analysis is important to social scientists, marketers and analysts because it provides demographic data directly from raw text or data [9]. Stylometric study is used to identify and authenticate the authorship of e-mail text messages [14]. The interest has been growing in applying stylometry to the content generation where the content is checked whether it is original or copied from others style. Shane Bergsma, Matt Post, David Yarowsky are evaluating stylometric techniques in the novel domain of scientific writing. Authors might also use these tools, e.g., to help ensure a consistent style in multi-authored papers , or to determine sections of a paper needing revision. The contributions of paper include, new Stylometric Tasks. They are predicting whether a paper is written:

(1) by a native or non-native speaker
(2) by a male or female, and
(3) in the style of a conference or workshop paper.

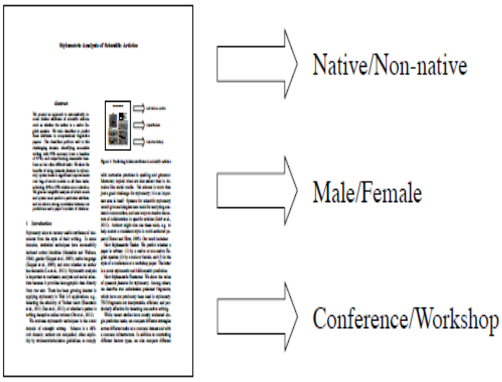

Figure 1 : Predicting hidden attributes in scientific articles [9]

Forensic linguistics has a sub-field that is forensic stylistics and the author identification can be done by applying stylistics. The stylistic is based on two premisses:

- Two writers do not write in the same pattern (having same mother-tongue).
- The writer itself does not write in the same pattern all the time.

The stylistic can be categories into two different approaches:

- Qualitative
- Quantitative

In the qualitative approach errors and personal behavior of the authors are assessed whereas in the quantitative approach focus on readily computable and countable language features, e.g. length of word, length of sentence, phrase length, frequency of vocabulary, distribution of words of different lengths [10]. Men and women technically speak the same language. There are lots of studies have been done to study the relationship between language use and gender. Gender identification problem can be treated as a binary classification problem in (a), i.e., given two classes {*male; female}*, assign an anonymous e-mail *e* to one of them according to the gender of the corresponding author:

e Є { *Class*1 if the author of *e* is male
e Є { *Class*2 if the author of *e* is female (a)

To test the binary hypothesis (a), set of features have to be selected that remain relatively constant for same gender written large number of e-mails. When the feature set has been selected, an *n*-dimensional vector represent a given e-mail, where *n* is the total number of features. A set of known pre-classified e-mails, a classifier (or model) can be built by classification techniques and the category of a new e-mail can be determined [11].

A. Gender: Male vs. Female

The author [9] have taken data of Bergsma and Lin (2006).This data has been widely used in conference resolution but never in stylometry. Each line in the data lists how often a noun co-occurs with male, female, neutral and plural pronouns. If the name has an aggregate count >30 and female probability >0.85, label as female; otherwise if the aggregate count is >30 and male probability >0.85, label male [9]. For the gender identification of the author of an e-mail is different from the other types of authorship identification problems. The e-mail length is usually not long as compared to other types of texts like books and novels. The e-mails style may vary according to the type or social status of recipients, for example, in business e-mails we follow formal style and in personal emails we follow informal style. Some special linguistic elements such as facial expressions often appear in e-mails. The format or the e-mails structure may vary among different users. Thus, specific email-based gender-differentiating feature sets must be considered along with traditional stylometric features [11]. Brennan and Greenstadt [12] explained that current authorship attribution algorithms are highly accurate in the non-adversarial case, but fail to attribute correct authorship when an author deliberately masks his writing style. The two forms of adversarial attacks were defined and tested: imitation and obfuscation.

In the imitation attack, authors hide their writing style by imitating another author. In the obfuscation attack, authors hide their writing style in a way that will not be recognized. Traditional authorship recognition methods perform less than random chance in attributing authorship in both cases. These results show that effective stylometry techniques need to recognize and adapt to deceptive writing. The author argued that some linguistic features change when people hide their writing style and by identifying those features, deceptive documents can be recognized. Deception requires additional cognitive effort to hide information, which often introduces subtle changes in human behavior. These behavioral changes affect verbal and written communication [13].

The task of identifying the author of a given text is author identification, therefore, it can be formulated as a typical classification problem, which depends on discriminant features to represent the style of an author [6]. Plagiarism detection [8] can be seen as complementary to stylometric authorship attribution: it attempts to detect common content between documents, even if the style may have been changed. Plagiarism is not always intentional or stealing some things from someone else; it can be unintentional or accidental and may comprise of self stealing.

4. Comparative Study

In this section we have compared the research based on scientific articles, e-mail author identification using stylometry on the basis of their result analysis. Shane Bergsma, Matt Post, David Yarowsky [9] have given the perforances analysis that they have considered *NativeL* (i.e Native vs Non-Native English Speaker) and *Venue (i.e Top-Tier Vs. workshop)*. For *NativeL*, they had made *Strict* rule (i.e. English name/country) and only plot papers marked as native. The papers which get the lowest *NativeL*-scores obtain fewer citations, but they soon level off (Figure 2(a)). They have analysed that many *junior* researchers at English universities are non-native speakers; early-career non-natives might receive fewer citations than well-known peers. The correlation between citations and *Venue*-scores is even stronger (Figure 2(b)).

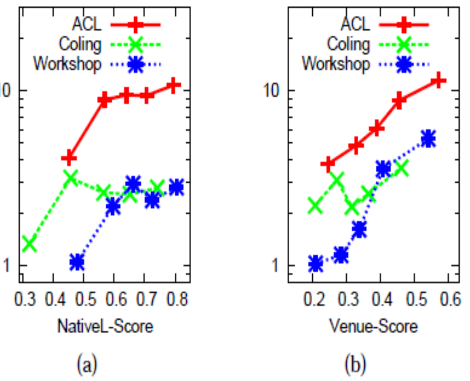

Fig. 2 Correlation between predictions (x-axis) and mean number of citations (y-axis, *log-scale*)[9].

The author [9] successfully calculated significant new tasks and techniques in the stylometric analysis of scientific article, included the novel resolution of publication venue based on paper style, and novel syntactic features based on tree substitution grammar fragments. In all above cases, there syntactic and stylistic features significantly improve over a bag-of-words (BOW) baseline, achieving 10%to 25% relative error reduction in all three major tasks. The author [14] created a program which was written in the C# programming language, and it is having a Graphical User Interface (GUI) to simplify the tasks of determining authorship by automating the identification process. For the determination of the authorship they involved collection of data, extraction of feature, and classification. Users helps the program to recognize authors by initially selecting a set of sample e-mails labeled with known authors (including author demographics) and subsequently selecting a set of sample e-mails by unknown authors for comparison. They consider fifty-five stylistic features. There were 12 participants and each participant created ten e-mails, which averaged one hundred and fifty (150) words, each on a distinct subject. In [14] they have used various techniques in pattern classification, such as Bayesian Theory, Decision Trees, Neural Networks or *k*-nearest neighbor (KNN). Their program uses the KNN algorithm which used to classify objects based on the basis of their similarities or distance metric. KNN classifiers are based on learning by analogy. On the basis of stylistics features they find out the result in the analytical manner. The dichotomy data for the Stylometry authentication experiments contained 1770 records for each subset of six subjects. Each subset was run against the other yielding 76.72% and 66.72% accuracy. The author [14] faces some difficulties and their future work is to extend the authentication task to identify patterns in frequently used misspelled and misused words.

5. Conclusion And Future Work

Through the overall discussion the paper we discussed first the basic behind the stylometry then later in the discussion move to the literature review where we have discussed about stylometry that can be used for the identification and authentication of the author in different fields like Author identification; detection of hoaxes, frauds, and deception in writing styles; gender identification from emails, plagiarism detection etc.. We have also analyse the result on the basis of the stylometric features for the scientific articles and email author identification. So in this manner, we just see that the stylometry can be used in many broad areas. A still lot of research has to be done in field of author identification but we have chosen to implement it for the security of email by identifying the author and with this the security of the email system will be improved.

6. References

[1] David I. Holmes, "*The Evolution of Stylometry in Humanities Scholarship*," Literary and Linguistic Computing 13/3: Pages: 111-117, 1998.

[2] T. C. Mendenhall, "*The Characteristic Curves of Composition*," Science 214, Pages : 237–246, 1887.

[3] C. Mascol, "*Curves of Pauline and Pseudo-Pauline Style I*," Unitarian Review, 30 Pages: 452–60, 1888; C. Mascol, "*Curves of Pauline and Pseudo-Pauline Style II*," Unitarian Review, 30 Pages: 539–46, 1888.

[4] L. A. Sherman, "*Analytics of Literature: A Manual for the Objective Study of English Prose and Poetry", Boston: Ginn, 1893.*

[5] Stefan Gruber, and Stuart Noven, "*Tool support for plagiarism detection in text documents*", Symposium on Applied Computing archive Proceedings of the 2005 ACM , Pages: 776 – 781, 2005.

[6] D. Pavelec, L. S. Oliveira, E.Justino, F. D. Nobre Neto, and L. V. Bastista, "*Compression and Stylometry for Author Identification*", Proceedings of International Joint Conference on Neural Networks, 2009.

[7] Burrows, J. F., "Computers and the Study of Literature",. In: C. S. Butler (ed.): Computers and Written Texts. Oxford: Blackwell, Pages: 167–204, 1992.

[8] H. Maurer, F. K., "*Plagiarism - a survey*", Journal of Universal Computer Science, vol. 12, no. 8 , Pages: 1050-1084, 2006.

[9] Shane Bergsma, Matt Post, David Yarowsky, "*Stylometric Analysis of Scientific Articles",* 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies , Pages: 327-337, 2012.

[10] Daniel Pavelec, Edson Justino, and Luis S.Oliveira, "*Author Identification using Stylometric Features*", Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial. Vol 11, No 36, Pages: 59-65, 2007.

[11] Na Cheng, Xiaoling Chen, R. Chandramouli, K. P. Subbalakshmi, "*Gender Identification from E-mails*", Computational Intelligence and Data Mining , Pages: 154-158, 2009.

[12] Michael Brennan, Rachel Greenstadt, "*Practical Attacks Against Authorship Recognition Techniques",* Innovative Applications of Artificial Intelligence (IAAI), 2009.

[13] Sadia Afroz, Michal Brennan and Rachel Greenstadt, "*Detecting Hoaxes, Frauds, and Deception in Writing Style Online*", IEEE Symposium on Security and Privacy, 2012 .

[14] K. Calix, M. Connors, D. Levy, H. Manzar, G. McCabe, and S. Westcott, "*Stylometry for E-mail Author Identification and Authentication*" Proceedings of CSIS Research Day, Pace University, May 2008.