Chapter 47
Authorship Analysis: Techniques and Challenges

Athira U.
LBS Center for Science and Technology, India

Sabu M. Thampi
IIITMK, India

ABSTRACT

Authorship Analysis is the process of examining documents to determine the stylistic details underlying the document and hence inferring about the characteristics of the author of document in order to attribute the authorship to a particular author or to confirm the authenticity of a claimed authorship. The popularity of online communications has paved way to the promotion of numerous fraudulent acts. These illegal activities can be curbed to an extent by identifying the source of the postings, which is made possible by finding the real authors of online documents. Applicability of authorship analysis in the field of forensic linguistics also gathers great importance today. The automation of, process aimed at analyzing the authorship of forensic documents, eases the linguists of the high manual effort spent in analyzing documents and is also advantageous in terms of its accuracy. Here we discuss about the existing methods that have been used so far to deal with automation of authorship analysis and the challenges faced by them.

INTRODUCTION

Authorship Analysis is the process of finding the author of a given document by analyzing the writing style followed in the document. It has been used to attribute authorships of many literary works in early days. The ancient example of authorship analysis was demonstrated by King Vikramadithya to find the greatest poet of his assembly, Kalidasa who was found absconding. The excellent comprehension of king in the poet's style enabled him to identify the source of certain lines written by delitescent Kalidasa.

Authorship analysis is mainly based on the assumption that each author is featured by a unique idiolect, which means a distinct and unique way of usage of language. The author's text reflects this characteristic and hence can be used to identify the authorship of a document (Rygl, 2013). The major

DOI: 10.4018/978-1-5225-1759-7.ch047

task behind the authorship analysis is to find out feature set representative of the idiolect of the author. Just like the unique finger print of a person, this feature set should be a unique write print capable of identifying an individual. This includes usage of function words, vocabulary richness, common terms used, frequency of n- lettered words and so on. As an example it is to be noted that the famous writer J.KRowling, the author of a series of seven fantasy novel Harry Potter, has a distinctive trait of usage of terms and spellings and also vocabulary set consists of terms characterizing her own unique style. Terms like "mudblood" is indeed the writer's own creativity. Similar is the case of Shakespeare who formulated terms suiting situations and characters. The same quality turned out to be a crucial evidence of identifying the so called "Unabomber", Ted Kaczynski the serial murderer. The convict wrote a letter to New York Times stating that he will detest from bombing if they were to publish his manifesto named Industrial Society and its future. The writing style followed in the manifesto was identified by his brother and sister- in- law which made the inquiry pretty easy. The phrase "cool-headed logician" was peculiar to Ted Kaczynski and was identified by his brother. Careful analysis of all literary works can reveal such hidden specialties of an author, which are to be captured correctly to represent the author characteristic. Authorship analysis aims at this type of analysis.

To begin with scientific studies, it has been used, to verify the authorship of certain plays of Shakespeare (Malone, 1787) . This analysis was a part of finding answer to the questioned closeness of works by Christopher Marlowe and William Shakespeare. Later several works were done on the same problem to prove the authorship of the disputed works. Similarly a statistical analysis was conducted by Scholar Kenneth Greyston and the statistician Gustav Herdan, in 1959–1960, to study the vocabulary usage of the contributors of Bible:The New Testament. A radical change in traditional approach of authorship analysis was observed in the methodologies followed by Mosteller and Wallace in 1964 to attribute authorship of the disputed federalist paper. Federalist papers were written as a part of swaying the population in NewYork to endorse the U.S Constitution in 1787-1788. Alexander Hamilton, John Jay and James Madison were the contributors of the paper comprising of 77 short essays. Later this together with 8 other essays pertaining the same area were compiled to form a book and was published in 1788. Out of 85 articles, 51 were written by Hamilton, 5 by John Jay, Madison wrote 14 papers and 3 papers were contributed by Madison and Hamilton. But there was an uncertainty regarding the authorship of 12 papers and it was doubted to be that of either Madison or Hamilton (Fung, 2003).Various analyses were conducted using Federalist paper as the data set. From then onwards methods for authorship attribution were rather computational than computer aided. The discussion so far dealt with the literary works. But the analysis of authorship finds its widespread applicability in several realms of real world scenario and hence procured a great scope of research.

As a part of curbing the spread of illegal postings including terrorism through cyber space, a research initiative named Dark Web Project has been undertaken by University of Arizona. They aim at establishing an immune intelligent system tolerant to the illicit content in the web as well as identifying such contents from web (Yang et al., 2012). The same strategy can be followed in identifying online fraudulence like impersonations and phishing sites.

The areas where finding the author of a document is very crucial, as in the case where language is used as evidence, make room for the need of automating the whole process of authorship analysis which is usually done manually by a linguist. It has been proved that suicide note classification using machine learning methods outperformed the decision of mental health professionals in predicting the forgery associated with suicide notes (Pestian et al, 2010). American Academy of Matrimonial Lawyers