A Systematic Review on Author Identification Methods

Sunil Digamberrao Kale, Smt. Kashibai Navale College of Engineering, Pune, India
Rajesh Shardanand Prasad, Computer Engineering Department, NBN Sinhgad School of Engineering (Savitribai Phule Pune University), Pune, India

ABSTRACT

Author Identification is a technique for identifying author of anonymous text. It has near about 130 year’s long history, started with the work by Mendenhall 1987. Applications of Author identification include plagiarism detection, detecting anonymous author, in forensics and so on. In this paper the authors outline features used for Author identification like vocabulary, syntactic and others. Researchers worked on various methods for Author identification they also outline this paper on types of Author Identification methods that include 1. Profile-based Approaches which includes Probabilistic Models, Compression Models, Common n-Grams (CNG) approach, 2. Instance-based Approaches which includes Vector Space Models, Similarity-based Models, Meta-learning Models and 3. Hybrid Approaches. At the end the authors conclude this paper with observations and future scope.

KEYWORDS

Attribution Method, Author Identification, Feature Set, Text Features, Text Mining

INTRODUCTION

Author identification is a technique of identifying author of a given document. Author identification method includes extraction of features from a given text data such as word length, vocabulary richness, use of digits, Letter Trigrams, Words, Letter Bigrams, Function Words, POS Bigram, POS tag, Letters, POS Trigrams, Prepositions, Word length, Pronouns, Conjunctions and so on. Statistical computations are applied on these feature set and compared on both training and test data. Process of identifying author is to provide training data as text written by different authors suppose 10 authors and 10 documents of each author (In total 100 text files) then identify feature set like Word Count, length of Statement, Richness of Word, Special Characters, Number of times use of particular Word, Punctuation, Conjunctions, Pronouns and so on. Next step is based on this features classify authors, then give input text i.e. test data that should belong to one of the authors used in training set and not same text data as which was used in training data, then with statistical computation calculate feature count of each feature of input text compare with feature count of authors used in training data where match is found that one is author for input text. The main idea behind statistically is computationally-supported authorship attribution that is measured by some textual features we can distinguish between texts written by different authors. Figure 1 shows general framework of Author Identification where Author-T1 to Author-Tn represents training data and Author-2 is identified author of unknown author i.e. anonymous text document by comparing results of training author’s data.

The first tries to quantify the writing in 19th century, by Mendenhall (1887) on the plays of Shakespeare followed by statistical studies. (Mosteller & Wallace, 1964) in his work used data set
as federalist paper these are essays of newspaper published in 1987 & 1988 by John Jay, Alexander Hamiltonian & James Madison. Total 85 essays out of those 5 essays wrote by John, 51 by Alexander and 14 by James. However, 3 are written jointly by Alexander and James. 12 out of 85 essays are disputed claimed by Alexander and James. Mosteller & Wallace used method. First they have used identifying feature as synonyms pairs but it not worked due to insufficient synonyms pairs then they have used 30 function words and that worked. This was the birth of statistical analysis they used probabilities and Bayesian analysis. Bayesian Classification based on conditional probabilities. (Halder, 2014) introduced a new concept of Bayesian decision theoretic rough set.

The organization of this paper is as follows. First we outline on features proposed for Author identification then proposed methods for Author identification at the end we conclude this paper with observation and future scope.

FEATURES USED FOR AUTHOR IDENTIFICATION

Vocabulary
The easiest way to confirm or refute authorship is simply to look for something that completely settles the authorship question. As a good example of this, the vocabulary of the Beale cipher (Kruh, 1982; Kruh, 1988) would seem to be conclusive proof of a hoax. The story is simple. In 1822, a man named Beale purportedly buried some treasure, leaving behind him a letter and a coded set of directions to find the treasure. This story was published in the 1860s, when the man to whom the letter was entrusted decided to go public. But did it happen that way?

One sentence of the letter of 1822 reads “Keeping well together they followed their trail for two weeks or more, securing many, and stampeding the rest.” Unfortunately, the word “stampede” is first attested by the Oxford English Dictionary as appearing in 1844; the earliest variant spelling (“stompado”) dates back to 1826, four years after the date on the letter. Similarly, the 1822 letter speaks of “improvised” tools, a meaning that the word “improvise” would not take until the 1850s. The word “appliance,” also appearing in the letter, had been used by Shakespeare, but was considered obsolete and outdated until it became popular again in the early 1860s. The conclusion is almost inescapable that the “Beale letters” are in fact late 19th century forgeries. One would be equally suspicious of a Hitler diary that mentioned the Hiroshima bombing or the construction of the Berlin wall. It is clear, then, that the individual words an author uses can be strong cues to his or her identity. The vocabulary labels the document as written at the time when the vocabulary existed, and most likely when it was current. Specific words can label the author by group identity — as hinted at in an earlier section, an author who writes of “colour,” “honour,” and “ironmongers” is likely to be British. Similarly, one who writes of sitting on a “chesterfield” is not only presumptively Canadian, but an older Canadian