Authorship Analysis Studies: A Survey

Sara El Manar El Bouanani  
ENSIAS - Mohammed V Souissi University  
Mohammed Ben Abdallah Regragui, Madinat Al Irfane, BP 713, Agdal Rabat, Morocco

Ismail Kassou  
ENSIAS - Mohammed V Souissi University  
Mohammed Ben Abdallah Regragui, Madinat Al Irfane, BP 713, Agdal Rabat, Morocco

ABSTRACT

The objective in this paper is to provide a review of the different studies done on authorship analysis. Focus is on outlining the Stylometric features that allow distinguishing between authors and on listing the diverse techniques used to classify an author’s texts.

General Terms
Authorship analysis

Keywords
Authorship characterization, authorship attribution, similarity detection, Stylometric features, probabilistic models, compression models, Machine learning classifiers, clustering algorithms, inter-textual distance.

1. INTRODUCTION

Authorship analysis is the process of examining the characteristics of a piece of work in order to draw conclusions on its authorship. Authorship analysis has its roots in a linguistic research area called stylometry, which refers to statistical analysis of literally style [26].

Authorship analysis measures some textual features and avoids distinguishing between texts written by different authors. First studies go back to 19th century with the study of Shakespeare’s plays on 1887 [49] followed by statistical studies in the first half of the 20th century ((58), [59], [61]). The study of the Federalist Papers [50] on 1964 was considered the most influential work in authorship attribution [56].

Authorship analysis studies can be classified into three categories ([11], [24] and [26]):

- **Authorship attribution or identification** determines the likelihood of a particular author having written a piece of work by examining other works produced by that author.

- **Authorship profiling or characterization** determines the author’s profile or the characteristics of the author that produced a given piece of work. These characteristics include gender, educational background, cultural background and language familiarity.

- **Similarity detection** compares multiple pieces of work and determines whether or not they are produced by a single author without necessarily identifying the author. Similarity is often used in the context of plagiarism detection which involves the complete or partial replication of a piece of work with or without permission of the original author.

Authorship analysis has been used in a diverse number of application areas. Many previous authorship studies focused on analyzing texts in the literature ([18], [19], [33], [40] and [50]), program codes ([27], [37] and [45]) and online messages ([11], [8], [11], [22] and [25]). However, with the growth of the web application and social networks, studies in the last decade focus on analyzing online messages (e-mails, blogs, forum…) rather than literary texts [56].

Therefore, the purpose in this paper is to present a survey of the studies and the techniques used in the field of authorship analysis. The paper is organized as follows. In Section 2, different areas of authorship analysis are described. Section 3 outlines the Stylometric features which are used to distinguish a text from another. Section 4 provides a survey of the different techniques used to detect authors. In Section 5, a brief conclusion is given.

2. AUTHORSHIP ANALYSIS

2.1 Authorship attribution

Authorship attribution is particularly concerned with the identification of the real author of a disputed anonymous document. In the literature, authorship identification is considered as a text categorization or text classification problem. The process starts by data cleaning followed by feature extraction and normalization. Each suspected document is converted into a feature vector [42]; the suspect represents the class label. Feature values are calculated by using Stylometric features. The extracted features are classified into two groups: training and testing sets. The training set is used to develop a classification model whereas the testing set is used to validate the developed model by assuming the class labels are not known. Common classifiers include decision trees, neural networks and Support Vector Machine [42].

Authorship attribution studies differ in terms of the Stylometric features used and the type of classifiers employed. References [30] and [47] describe two approaches which attempt to mine e-mail authorship for the purpose of computer forensics. Authors extract various e-mail document features including linguistic features, header features, linguistic patterns and structural characteristics. All these features are used with the Support Vector Machine (SVM) learning algorithm to attribute authorship of e-mail messages to an author.

Reference [26] develops a framework for authorship identification in online messages to deal with the identity-tracing problem. In this framework, four types of writing style features (lexical, syntactic, structural and content-specific features) are extracted from English and Chinese online-newsgroup messages. Comparison has been made between three classification techniques: decision tree, SVM and back-propagation neural networks. Experimental results showed that this framework is able to identify authors with a satisfactory accuracy of 70 to 95% and the SVM classifier outperformed the two others.

Reference [60] uses only function words and applies five classifiers (Naïve Bayesian, Bayesian networks, Nearest-neighbor method, Decision Trees, SVM). The data analyzed is a collection of newswire articles from the AP (Associated Press) sub-collection.

2.2 Authorship Characterization

Authorship characterization is used to detect sociolinguistic attributes like gender, age, occupation and educational level of the potential author of an anonymous document [42].
References ([9], [10], [11], [14] and [15]) studied the effects of gender attributes on authorship analysis. Other studies discussed the educational level, age and language background ([9], [50]). Reference [15] collected information about gender, age and occupation of the writer of an anonymous chat segment.

2.3 Authorship Verification or Similarity Detection

Studies consider the problem of authorship verification as a similarity detection problem: to determine whether two texts are produced by the same person without knowing the real author of the document [42].

Reference [51] proposes a new algorithm to identify when two aliases belong to the same individual, while preserving privacy. The technique has been successfully applied to postings of different bulletin boards, achieving more than 90% accuracy.

References [3] and [4] present a novel technique called writeprints for authorship identification and similarity detection. Authors used in the experimentation extended feature list, including idiosyncratic features. Authors take an anonymous entity, compare it with all other entities, and then calculate a score. If the score is above a certain predefined value, the entity is clustered with the matched entity.

Reference [57] proposes an approach called linguistic profiling. In this study [57] proposed some distance and scoring functions for creating profiles for a group of example data. The average feature counts for each author was compared with a general stylistic profile built from the training samples of widely selected authors. The study focused on detecting similarity between student essays for plagiarism and identity theft.

3. STYLOMETRIC FEATURES

Stylistics or the study of Stylometric features shows that individuals can be identified by their relatively consistent writing styles. The writing style of an individual is defined by the terms used, the selection of special characters, and the composition of sentences... Studies in literature show that there are no such features set optimized and applicable to all people and to all domains [41]. Four types of Stylometric features are defined: lexical, syntactic, structural and content-specific features. In this section, a description of each type of these features is given.

3.1 Lexical features

A text can be viewed as a sequence of tokens grouped into sentences. A token can be a word, a number or a punctuation mark. Earlier studies in authorship attribution were based on simple measures such as sentence length counts and word length counts. The advantage of these features is that they can be applied to any corpus in any language and with no additional requirements except the availability of a tokenizer [56].

Lexical features are used to learn about the preferred use of characters and words of an individual. As example, these features include frequency of individual alphabets, frequency of special characters, total number of upper case letters, capital letters used in the beginning of sentences, average number of characters per word, average number of characters per sentence. A text can also be viewed as a sequence of characters. Various character-level measures can be defined, including alphabetic characters count, digit characters count, uppercase and lowercase characters count, letter frequencies, punctuation marks count [56].

Vocabulary richness functions quantify the diversity of the vocabulary of a text. Some examples of this measure are the ratio V/N, (V is the size of the vocabulary and N is the total number of tokens of the text), Yule’s K measure, number of hapax legomena (words occurring once), number of hapax dislegomena (words occurring twice) [30]. Unfortunately, the vocabulary size heavily depends on text-length [56]. Various functions have been proposed to achieve stability over text-length, including Yule’s K measure, Simpson’s D measure, Sichel’s S measure, Brunet’s W measure, Honore’s R measure.

Another method to define a lexical feature set is to extract the most frequent words in the corpus. Reference [31] uses the 250 most frequent words, [54] extracts the 1000 most frequent words and [12] uses words that appear at least twice in the corpus.

From another point of view, [44] proposed various writing error measures to capture the idiosyncrasies of an author’s style. To that end, it defined a set of spelling errors (letter omissions and insertions) and formatting errors (all caps words) and it proposed a methodology to extract such measures automatically using a spell checker.

3.2 Syntactic features

Reference [1] defines syntactic features as the patterns used to form sentences. This category of features consists of the tools used to structure sentences. These include punctuation and function words. Function words are the common words (articles, preposition, pronouns...) like while, upon, though, where, your. Studies based on function words are listed in ([1], [11], [44] and [60]). Authors use a set of function words varying between 150 and 675 functions.

3.3 Structural features

Structural features are helpful in learning about how an individual organizes the structure of his documents. For instance, how sentences are organized within paragraphs and paragraphs within documents. Structural features were first suggested by [30] for e-mail authorship attribution. In addition to the general structural features, authors in [30] used specific features to e-mails such as the presence/absence of greetings and farewell remarks and their position within the e-mail body.

3.4 Content-specific features

Content-specific features are used to characterize certain activities, discussion forums or interest groups by a few keywords or terms [41]. Authors in [26] manually observe, analyze historical messages and identify 11 key words as content-specific features particularly for English “for-sale” online messages (Obo, sale, windows, software, Microsoft …). Based on the discussion above, a listing of the most useful Stylometric features and an overview of previous studies dealing with these features are proposed in table 1 and 2.

<table>
<thead>
<tr>
<th>Table 1: Stylometric features</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lexical features (F1)</strong></td>
</tr>
<tr>
<td><strong>Character-based features</strong></td>
</tr>
<tr>
<td>Characters count (C)</td>
</tr>
<tr>
<td>Total number of alphabetic characters/C</td>
</tr>
<tr>
<td>Total number of upper-case characters/C</td>
</tr>
<tr>
<td>Total number of digit characters/C</td>
</tr>
<tr>
<td>Total number of white-space characters/C</td>
</tr>
<tr>
<td>Total number of tab spaces/C</td>
</tr>
<tr>
<td>Frequency of letters (26 features) A-Z</td>
</tr>
<tr>
<td>Frequency of special characters ~, @, #, $, %, ^, &amp;, *, _, =, +, &gt;, &lt;, [ ], { }, /, \</td>
</tr>
</tbody>
</table>
4. AUTHORSHIP DETECTION TECHNIQUES

In literature, authors consider two sets of texts in every authorship detection problem. The first one is a set of candidate authors; texts of known authors which called the training corpus. The second one is a set of texts of unknown authors called the test corpus. Each one of these texts should be attributed to a candidate author [56].

One way to handle the available training texts per author is to concatenate them in one single text file. This file is used to extract the properties of the author’s style. Text of an unknown author is, then, compared with each author’s file and the most likely author is estimated based on a distance measure. As a result, the differences between the training texts by the same author are discarded [56]. In the literature, this first approach is realized by using Probabilistic models and compression models.

Another family of approaches requires multiple training text samples per author in order to develop an accurate attribution model. This means that each training text is individually represented as a separate instance of authorial style [56]. In the literature, the second approach is adopted by using Machine learning classifiers and clustering algorithms and inter-textual distance.

4.1 Probabilistic Models

The first approach in author identification presented in this work is the probabilistic models or the Bayesian classifier based on bayes theorem ([12], [38], [52] and [60]).

This method is based on the assumption that the occurrences of the features are mutually independent. Under this assumption, given the set of features \{a_1, \ldots, a_n\} extracted from a document and an author \(v\), we wish to compute

\[
P(v | a_1, \ldots, a_n) = \frac{P(v) P(a_1, \ldots, a_n | v)}{P(a_1, \ldots, a_n)}
\]

Where \(P(a_1, \ldots, a_n)\) is assumed to be uniform and \(n\) is fixed. Thus we can attribute the document to be classified by computing

\[
P(a_1, \ldots, a_n | v) = \Pi_1 P(a_i | v)
\]

Using Bayes theorem, a naive Bayesian classifier can be, thus, written as:

\[
v = \arg \max_{v \in V} P(v) \Pi_1 P(a_i | v)
\]

Where \(P(v)\) can be estimated by measuring the frequency with which author \(v\) occurs in the training data.

Reference [60] uses 365 function words and applies naive bayes and bayes network to identify the author of an unattributed document. Data analyzed is a collection of newswire articles from the AP (Associated Press) sub collection. Each document is represented as a vector with 365 dimensions (every dimension is a function word). The magnitude of each feature is calculated from the normalized frequency of the word in that document. Experiments have proven that this classification using probabilistic model gives a high accuracy varying from 78% to 90.46%.

An extension of the naive Bayes algorithm augmented with statistical language models was proposed by [52] and achieved
high performance in authorship attribution experiments. In comparison to standard naive Bayes classifiers, the approach of [52] allows local Markov chain dependencies in the observed variables to capture contextual information. Reference [21] addresses the problem of classification of articles of ambiguous authorship to the articles written by contemporary Tamil scholars by using probabilistic neural network.

4.2 Compression models
Initially, all the available texts for the i-th author are first concatenated to form a big file xa and a compression algorithm is called to produce a compressed file C(xa). Then, the unseen text x is added to each text xa and the compression algorithm is called again for each C(xa+x). The difference in bit-size of the compressed files d(x,xa)=C(xa+x)–C(xa) indicates the similarity of the unseen text with each candidate author.

Essentially, this difference calculates the cross-entropy between the two texts. Several off-the-shelf compression algorithms have been tested with this approach including RAR, LZW, GZIP, BZIP2, 7ZIP... and in most of the cases RAR found to be the most accurate [56].

Reference [7] defines the Normalized Compressor Distance (NCD). Given a compressor C and two documents x, y, this distance is defined as:

\[ NCD(x, y) = \frac{C(y) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}} \]

Where C(x), C(y) and C(xy), are the bit-wise sizes of the result sequences when using C to compress x, y and the concatenation of x and y, respectively. NCD assesses the similarity between a pair of documents by measuring the improvement achieved by compressing an information-rich document using the information found in the other document. Authors in [7] analyze patterns in the contributions of Serial Sharers on the web. They examine their individual contributions and propose a method for detecting their pages in large and diverse collections of pages by using compression distance.

Reference [43] compares different compression models for authorship attribution. To this end, three different types of compressors, namely GZip, BZip and PPM, along with two different similarity measures were considered. Results revealed that compression models are a good alternative for authorship attribution surpassing pattern recognition systems based on classifiers and feature extraction.

4.3 Machine learning classifiers and Clustering algorithms
The use of machine learning classifiers and clustering algorithms marked an important turning point in authorship attribution studies. The application of such methods is straightforward: training texts are represented as labeled numerical vectors and learning methods are used to find boundaries between classes (authors) that minimize some classification loss function [14].

To predict the performance of a particular algorithm, accuracy measure, precision and recall are used. They are defined as:

\[ \text{Accuracy} = \frac{\text{Number of messages whose author was correctly identified}}{\text{Total number of messages}} \]

\[ \text{Precision} = \frac{\text{Number of messages correctly assigned to the author}}{\text{Total number of messages assigned to the author}} \]

\[ \text{Recall} = \frac{\text{Number of messages correctly assigned to the author}}{\text{Total number of messages written by the author}} \]

As an example of using clustering algorithms, [41] tested the three algorithms K-means, EM and bisecting K-means on a set of e-mails. Texts are converted into a vector of Stylometric features and clustering algorithms are applied to bring together texts written by the same author in a cluster.

References [26] and [24] develop a framework for authorship identification in English and Chinese online messages to address the identity-tracing problem. In this framework, four type of writing style features (lexical, syntactic, structural and content-specific features) are extracted. Three classification techniques are compared: decision tree, SVM and back-propagation neural networks. Experimental results showed that this framework can identify authors with satisfactory accuracy of 70 to 95%. SVM outperformed the other two classifiers. Reference [1] explores the problem of analyzing extremist group web forum messages and proposed a framework for analyzing online messages in Arabic and English. Two machines learning classifiers are used: C4.5 decision tree algorithm and SVM. Experimental results show a high accuracy of 94, 83% for Arabic data and 97% for English one. Reference [8] describes an investigation into e-mail content mining for authorship attribution, for the purpose of forensic investigation. Authors focus on the ability to discriminate between authors for the case of both aggregated e-mail topics as well as across different email topics. An extended set of e-mail document features including structural characteristics and linguistic patterns were derived and SVM learning algorithm was used for mining the e-mail content. Finally, [9] describes an investigation of authorship gender attribution mining from e-mail text documents. Authors used an extended set of predominantly topic content-free e-mail document features such as style markers, structural characteristics and gender-preferential language features together with a Support Vector Machine learning algorithm.

4.4 Inter-textual Distance
The main idea of distance approaches is: if the vocabulary used in two texts is similar, both texts are closer and it is possible that there were written by the same person. The result “two texts are closer” is obtained by measuring the distance between them.

Reference [53] explains the concept of using inter-textual distance in authorship detection and details the most useful distances: Delta measure, chi-square distance and Kullback-Leibler Divergence.

Reference [19] suggests accounting for the most frequent word types (and particularly function words) without taking punctuation marks or numbers into account. Author suggests considering from 40 to 150 the most frequently occurring word types, with 150 words obtaining the best results by applying the delta measure based on standardized scores. This measure was also tested by Hoover in [40].

Reference [53] evaluates the 3 distances in two experiments. The first was based on 5408 newspaper articles (Glasgow Herald) written in English by 20 distinct authors and the second on 4326 newspaper articles (La Stampa) written in Italian by 20 distinct authors. The experiments clearly revealed that Delta measure outperforms the two other distances.

5. CONCLUSION
In this paper, different feature types and methods proposed in the field of authorship analysis were surveyed. This field covers several text domains (newspaper articles, online forum messages and blogs, emails, literary works) and several languages (French, English, Chinese...). The methods used are
the same regardless of the language of the texts and especially those based on Stylometric features can be easily applied to any language.

Various studies have clearly shown that the result of an authorship attribution method can be affected by parameters such as training corpus size, test corpus size, length of the texts (certain methods work effectively in the case of long texts but not well on short or very short texts), number of candidate authors and distribution of the training corpus over the authors.

Other factors that should be controlled in an evaluation corpus include age, nationality and gender. In addition, all the texts per author should be written in the same period because a writing style can change over time and ideally, all the texts of the training corpus deal with exactly the same topic for all candidate authors. An attribution method should also be tested on a variety of text genres (newspaper, blogs…) to reveal if it can be applied to different texts domains or simply to a specific one.

At the end, Table 3 presents the most important studies found in literature. For each one, the corpus, its language and domain, the features types used and the method(s) tested are listed.

Table 3: The most important studies on authorship analysis

**F1**: Lexical features; **F2**: Syntactic features; **F3**: Structural features; **F4**: Content-specific features; **A1**: Authorship identification; **A2**: Authorship characterization; **A3**: Similarity detection

<table>
<thead>
<tr>
<th>Year</th>
<th>Reference</th>
<th>Authorship detection field</th>
<th>Corpus</th>
<th>Domain</th>
<th>Features</th>
<th>Language</th>
<th>Techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999</td>
<td>[34]</td>
<td>A1</td>
<td>Greek weekly newspaper TOBHMA</td>
<td>N/A</td>
<td>F2, F3</td>
<td>Greek</td>
<td>PCA</td>
</tr>
<tr>
<td>2001</td>
<td>[36]</td>
<td>A1</td>
<td>Modern Greek newspaper</td>
<td>N/A</td>
<td>Greek</td>
<td>MDA</td>
<td></td>
</tr>
<tr>
<td>2001</td>
<td>[35]</td>
<td>A1, A2</td>
<td>Texts (journalistic, scientific…)</td>
<td>N/A</td>
<td>F1, F2</td>
<td>Greek</td>
<td>MDA</td>
</tr>
<tr>
<td>2002</td>
<td>[9]</td>
<td>A2</td>
<td>4369 e-mails 325 authors</td>
<td>Computer forensics</td>
<td>F1, F2, F3</td>
<td>SVM</td>
<td></td>
</tr>
<tr>
<td>2002</td>
<td>[16]</td>
<td>A1</td>
<td>72 Literary texts (fiction, argument and description)</td>
<td>literary</td>
<td>F1, F2</td>
<td>Dutch</td>
<td>PCA LDA</td>
</tr>
<tr>
<td>2003</td>
<td>[1]</td>
<td>A2</td>
<td>Electronic message</td>
<td>N/A</td>
<td>F1, F2, F3</td>
<td>Exponentiated Gradient algorithm</td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td>[51]</td>
<td>A3</td>
<td>Posting messages <a href="http://www.courttv.com">www.courttv.com</a></td>
<td>N/A</td>
<td>F1, F2, F3 - correlation of posting times -analysis of</td>
<td>KLD distance</td>
<td></td>
</tr>
<tr>
<td>Year</td>
<td>Reference</td>
<td>A1</td>
<td>A2</td>
<td>Domain</td>
<td>Language</td>
<td>Model</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>-----------</td>
<td>----</td>
<td>----</td>
<td>--------</td>
<td>----------</td>
<td>-------</td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td>[26]</td>
<td>A1</td>
<td>-</td>
<td>Online messages</td>
<td>F1, F2, F3, F4</td>
<td>SVM</td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>[55]</td>
<td>A3</td>
<td>-</td>
<td>N/A</td>
<td>N/A</td>
<td>SVM</td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td>[41]</td>
<td>A1</td>
<td>-</td>
<td>forensic investigation</td>
<td>F1, F2, F3, F4</td>
<td>English</td>
<td></td>
</tr>
</tbody>
</table>
6. REFERENCES


In theory, authorship sounds straightforward, but in practice it often causes headaches. While preparing these guidelines, we heard about several cases. Design, or acquisition of data, or analysis and interpretation of data; (2) drafting the article or revising it critically for important intellectual content; and (3) final approval of the version to be published. Conditions (1), (2), and (3) must all be met. Even before a study is finished, you should have some idea of the publications that might come out of it, such as a conference abstract, the full paper, then some supplementary papers, and who is likely to be most involved in these. Goodman NW. Survey of fulfillment of criteria for authorship in published medical research.