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 ([1], [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 ([1], [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 (Naive 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 1: Stylometric features

Lexical features (F1)
Character-based features
Characters count (C)
Total number of alphabetic characters/C
Total number of upper-case characters/C
Total number of digit characters/C
Total number of white-space characters/C
Total number of tab spaces/C
Frequency of letters (26 features) A–Z
Frequency of special characters ~ , @, #, $, %, ^, &, *, -, _, = ,+, >, <, [,], {, }, /, \, |

Word-based features
- Total number of words (M)
- Total number of short words (less than four characters)/M e.g., and, or
- Total number of characters in words/C
- Average word length
- Average sentence length in terms of character
- Average sentence length in terms of word
- Total different words/M
- Hapax legomena
- Hapax dislegomena
- Yule's K measure
- Simpson's D measure
- Sichel's S measure
- Brunet's W measure
- Honore's R measure

Syntactic Features (F2)
- Frequency of punctuations ",", ".", "?", "!", ";", ":", "'", "(", ")"
- Frequency of function words

Structural Features (F3)
- Total number of lines
- Total number of sentences
- Total number of paragraphs
- Number of sentences per paragraph
- Number of characters per paragraph
- Number of words per paragraph
- Has a greeting
- Has a separator between paragraphs
- Use e-mail as signature
- Use telephone as signature
- Use URL as signature

Content-specific Features (F4)
- Frequency of content specific keywords

Table 2: Previous studies using Stylometric features

References	Features			
	F1	F2	F3	F4
[39] - 1992	×			
[34] - 1999		×	×	
[8] - 2001	×	×	×	
[35] - 2001	×	×	×	
[9] - 2002	×	×	×	
[10] - 2002		×		
[16] - 2002	×	×		
[17] - 2002		×		
[11] - 2003	×	×	×	
[24] - 2003		×	×	×
[6] - 2003	×		×	
[51] - 2004	×	×	×	
[1] - 2005	×	×	×	×
[13] - 2005		×		
[22] - 2005	×	×	×	
[60] - 2005		×		
[25] - 2006	×	×	×	×
[2] - 2006	×	×	×	×
[5] – 2006	×	×		
[26] - 2006	×	×	×	×
[28] - 2008	×	×		×
[41] – 2010	×	×	×	×
[28] – 2010	×	×	×	×
[21] – 2012		×		
[32] – 2012	×	×		

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 {a1. . . an} extracted from a document and an author v, we wish to compute

$$P(v|a_1,\ldots,a_n) = \frac{P(v)\cdot P(a_1,\ldots,a_n|v)}{P(a_1,\ldots,a_n)}$$

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

$$P(a_1,\ldots,a_n|v) = \Pi_i\, P(a_i|v)$$

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

$$v = \text{argmax}_{v\in V}\; P(v)\,\Pi_i\, P(a_i|v)$$

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 naïve 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 naïve Bayes algorithm augmented with statistical language models was proposed by [52] and achieved

high performance in authorship attribution experiments. In comparison to standard naïve 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-wise 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:

$$NCD(x,y)=\frac{C(xy)-\min\{C(x),C(y)\}}{\max\{C(x),C(y)\}}$$

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*

Year	Reference	Authorship detection field	Corpus	Domain	Features	language	Techniques
1992	[39]	A1	Book of Mormon	literary	F1		PCA
1999	[34]	A1	Greek weekly newspaper TOBHMA	N/A	F2, F3	Greek	PCA
2001	[46]	A1	Moliere and Corneille works	Literary	vocabulary	French	Intertextual distance
2001	[8]	A1	e-mails	Computer forensics	F1, F2, F3		SVM
2001	[36]	A1	Modern Greek newspaper	N/A		Greek	MDA
2001	[35]	A1 A2	Texts (journalistic, scientific…)	N/A	F1, F2	Greek	MDA
2002	[9]	A2	4369 e-mails 325 authors	Computer forensics	F1, F2, F3		SVM
2002	[10]	A2	566 documents http://shekel.jct.ac.il/~argamon/gender-style	N/A	F2	British English	Exponentiated Gradient algorithm
2002	[16]	A1	72 Literary texts (fiction, argument and description)	literary	F1, F2	Dutch	PCA LDA
2002	[17]	A1	The Royal Book of Oz	Literary	F2	English	PCA
2003	[20]	A1	- Modern greek weekly newspaper TOBHMA -books from different authors	Literary	N-grams	English Chinese Greek	Dissimilarity measure
2003	[1]	A2	Electronic message	N/A	F1, F2, F3		Exponentiated Gradient algorithm
2003	[24]	A1	-e-mails -web messages	Cyber crime	F2, F3, F4	English Chinese	-Decision tree -Neural Network -SVM
2003	[6]	A1	hurriyet.com.tr		F1, F3	Turkish	Multilayer Perceptron algorithm
2004	[51]	A3	Posting messages www.courttv.com	N/A	F1, F2, F3 - correlation of posting times -analysis of		KLD distance

					signature files -clustering of misspellings -references to entities -expressed relationships -use of blank lines -use of HTML tags (fonts, colors, or links		
2004	[52]	A1 A2	-texts from modern Greek authors -Newsgroup		N-grams	Greek English Japanese Chinese	naive Bayes models
2005	[1]	A1	Forum messages	Cybercrime	F1, F2, F3, F4	English Arabic	-Decision tree -SVM
2005	[13]	A1	twenty novels	literary	F2	English	SVM
2005	[22]	A1	Chaski's Writing Database	digital crime	F1, F2, F3 -n-grams -part of speech		Discriminant function analysis (DFA)
2005	[60]	A1	newswire articles from the AP (Associated Press) sub collection		F2		-Naive Bayesian -Bayesian networks -Decision Trees
2006	[25]	A1	Online messages -misc.forsale.com -smth.org -mitbbs.net	Cybercrime	F1, F2, F3, F4	English Chinese	Genetic algorithms
2006	[2]	A1	Online forums -Yahoo group forum for Al-Aqsa Martyrs -Web site forum for the White Knights Ku Klux Klan	Cybercrime	F1, F2, F3, F4	English Arabic	-PCA - SVM
2006	[5]	A1	Biblical texts		F1, F2	Koine Greek	-MDA -trigram Markov method
2006	[26]	A1	Online messages		F1, F2, F3, F4		-Decision tree -Neural Network -SVM
2007	[7]	A3	10000 web pages 2201 authors	N/A	vocabulary		Compression distance
2008	[28]	A1	e-mails	Cybercrime	F1, F2, F3, F4		Frequent patterns
2009	[55]	A3	N/A	N/A	N-grams		Dissimilarity measure
2010	[41]	A1	e-mails	forensic investigation	F1, F2, F3, F4 common spelling mistakes and grammatical mistakes such as sentences containing incorrect	English	-Clustering algorithms (K-means, EM) - Frequent patterns
2010	[28]	A1	e-mails	Cyber crime	F1, F2, F3, F4	English	-Naive Bayes -Bayesian Network -SVM with Sequential Minimum Optimization -SVM with RBF

							kernel
2010	[48]	A1	Poetic work of Aragon (1917-1952)	literary	Vocabulary	French	Intertextual distance implemented in HYPERBASE
2011	[23]	A1	N/A	N/A	Character bi-gram and tri-gram	English	Dissimilarity measure
2012	[21]	A1	Literary works of three contemporary Tamil scholars	literary	F2	Indian	Probabilistic Neural Network
2012	[32]	A1	daily newspaper SABAH (www.sabah.com.tr)	N/A	F1, F2	Turkish	Similarity function

6. REFERENCES

[1] A. Abbasi, H. Chen 2005. Applying authorship analysis to extremist-group web forum messages. IEEE Intelligent Systems, 20(5), pp : 67-75.

[2] A. Abbasi, H. Chen 2006. Visualizing Authorship for Identification. ISI, LNCS 3975, pp : 60-71.

[3] A. Abbasi, H. Chen 2008. Writeprints: A Stylometric approach to identity-level identification and similarity detection in cyberspace. ACM Transactions on Information Systems, 26(2), pp : 1-29.

[4] A. Abbasi, H. Chen, J. Nunamaker 2008. Stylometric identification in electronic markets: Scalability and robustness. Journal of Management Information Systems, 5(1), pp : 49-78.

[5] D. Abbott, M.J. Berryman, S. Jain, T.J. Putnins, D.J. Signoriello 2006. Advanced text authorship detection methods and their application to biblical texts. The International Society for Optical Engineering, pp : 1-13.

[6] M. Amasyali, B. Diri 2003. Automatic Author Detection for Turkish Texts. ICANN2003

[7] E. Amitay, S. Yogev, E. Yom-Tov 2007. Serial Sharers: Detecting Split Identities of Web Authors. Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection. ACM SIGIR Amsterdam.

[8] A. Anderson, M. Corney, O. DeVel, G. Mohay 2001. Mining E-mail Content for Author Identification Forensics. SIGMOD Record, 30(4), pp : 55-64.

[9] A. Anderson, M. Corney, G. Mohay, O. DeVel 2002. Gender-preferential text mining of e-mail discourse. In ACSAC'02: Proc. of the 18th Annual Computer Security Applications Conference, Washington, DC, pp : 21-27.

[10] S. Argamon, M. Koppel, A.R. Shimoni 2002. Automatically categorizing written texts by author gender. Literary and Linguistic Computing, 17(4), pp : 401-412.

[11] S. Argamon, M. Saric, S. Stein 2003. Style mining of electronic messages for multiple authorship discrimination: First results. In Proceedings of the 9th ACM SIGKDD, pp : 475-480.

[12] S. Argamon, D. Fradkin, A. Genkin, D. Lewis, D. Madigan, L. Ye 2005. Author identification on the large scale. In Proceedings of CSNA-05.

[13] S. Argamon, S. Levitan 2005. Measuring the usefulness of function words for authorship attribution. Proceedings of the 2005 ACH/ALLC Conference

[14] S. Argamon, M. Koppel, J. Schler 2009. Computational methods in authorship attribution". J. Am. Soc. Inf. Sci. Technol., 60(1), pp : 9-26.

[15] C. Aykanat, B. B. Cambazoglu, F. Can, T. Kucukyilmaz 2008. Chat mining: predicting user and message attributes in computer-mediated communication. Information Processing and Management, 44(4), pp : 1448-1466.

[16] R. Baayen, A. Neijt, F. Tweedie, H. VanHalteren 2002. An experiment in authorship attribution. In proceedings of the 6th International Conference on the Statistical Analysis of Textual Data (JADT).

[17] G. Binongo, J. Nilo 2002. Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution. Conference of the Classification Society of North America.

[18] R. Bosch, J. Smith 1998. Separating hyper planes and the authorship of the disputed federalist papers. American Mathematical Monthly, 105(7), pp : 601-608.

[19] J. Burrows 2002. Delta: A measure of stylistic difference and a guide to likely authorship. Literary and Linguistic Computing, 17(3), pp : 267-287.

[20] N. Cerconey, V. Keselj, F. Peng, C. Thomas 2003. N-Gram based author profiles for authorship attribution, Pacific Association for Computational Linguistics.

[21] R. Chandrasekaran, G. Manimannan 2012. Use of Probabilistic Neural Network In The Classification Of Articles Of Ambiguous Authorship. International Journal of Engineering Research & Technology (IJERT),Vol. 1 Issue 7.

[22] C. Chaski 2005. Who's at the keyboard? Authorship attribution in digital evidence investigations. International Journal of Digital Evidence. Vol. 4, Issue 1.

[23] M. Chaurasia, H.F. Hassan 2011. Author Assertion of Furtive Write Print Using Character N-Grams. International Conference on Future Information Technology IPCSIT vol.13.

[24] H. Chen, Z. Huang, Y. Qin, R. Zheng 2003. Authorship Analysis in Cybercrime Investigation (Eds.): ISI 2003, LNCS 2665, pp : 59-73.

[25] H. Chen, J. Li, R. Zheng. 2006. From fingerprint to writeprint. Communications of the ACM - April 2006/Vol. 49, No.4.

[26] H. Chen, Z. Huang, J. Li, R. Zheng 2006. A framework for authorship Identification of Online Messages: writing-Style features and classification Techniques. JASIST, pp : 378-393.

[27] R. Cook, W.P. Oman. Programming style authorship analysis. In the proceeding of the 17th annual ACM computer Science Conference, pp : 320-326.

[28] M. Debbabi, B.C.M. Fung, R. Hadjidj, F. Iqbal 2008. A novel approach of mining write-prints for authorship attribution in e-mail forensics. digital investigation 5, pp : 42-51.

[29] M. Debbabi, B.C.M. Fung, F. Iqbal, L.A. Khan 2010. E-mail Authorship Verification for Forensic Investigation. SAC'10 March 22-26, 2010, Sierre, Switzerland. Copyright 2010 ACM 978-1-60558-638-0/10/03.

[30] O. DeVel, 2000. Mining e-mail authorship. In Proceeding of the Workshop on text mining in ACM international conference on knowledge discovery and data mining.

[31] E. Dokow, M. Koppel, J. Schler, 2007. Measuring differentiability: Unmasking pseudonymous authors. Journal of Machine Learning Research, 8, pp : 1261-1276.

[32] E. Ekinci, H. Takçı 2012. Character Level Authorship Attribution for Turkish Text Documents, TOJSAT: The Online Journal of Science and Technology- July 2012, Vol.2, Issue 3.

[33] W. Elliot, R. Valenza 1991. Was the Earl of Oxford the true Shakespeare? Notes and Queries, 38, pp : 501-506.

[34] N. Fakotakis, G. Kokkinakis, E. Stamatatos 1999. Automatic Authorship Attribution. Proceedings of EACL '99.

[35] N. Fakotakis, G. Kokkinakis, E. Stamatatos 2001. Automatic Text categorization in Terms of Genre and author. Computational Linguistics Vol.26, Issue 4.

[36] N. Fakotakis, G. Kokkinakis, E. Stamatatos 2001. Computer-Based Authorship Attribution without Lexical Measures. Computers and the Humanities 35, pp : 193-214.

[37] G. Frantzeskou, S. Gritzalis, S. Katsikas, E. Stamatatos 2006. Effective identification of source code authors using byte-level information. In Proceedings of the 28th International Conference on Software Engineering, pp : 893-896.

[38] S. Guenter, C. Sanderson 2006. Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation. In Proceedings of the International Conference on Empirical Methods in Natural Language Engineering, pp : 482-491.

[39] D. Holmes 1992. A Stylometric Analysis of Mormon Scripture and Related Texts. Journal of the Royal Statistical Society. Series A (Statistics in Society), Vol. 155, No. 1, pp : 91-120.

[40] D. Hoover 2004. Testing Burrows' Delta. Literary and Linguistic Computing, 19(4), pp : 453-475.

[41] F. Iqbal 2010. Mining writeprints from anonymous e-mails for forensic investigation. Digit Investig, doi:10.1016/j.diin.2010.03.003.

[42] F. Iqbal 2011. Messaging Forensic Framework for Cybercrime Investigation. A Thesis in the Department of Computer Science and Software Engineering - Concordia University Montréal, Canada.

[43] E. Justino, L.S. Oliveira, W. Oliveira Jr 2013. Comparing compression models for authorship attribution. Forensic Science International 228, pp : 100-104.

[44] M. Koppel, J. Schler 2003. Exploiting stylistic idiosyncrasies for authorship attribution. In Proceedings of IJCAI'03.

[45] I. Krsul, H.E. Spafford. Authorship analysis: identifying the author of a program. Computer security 16, 3, pp : 233-257.

[46] C. Labbe, D. Labbe 2001. Inter-textual distance and authorship attribution Corneille and Moliere, Journal of Quantitative Linguistics. 8-3, December 2001, pp : 213-231.

[47] M. Lai, Y. Li, J. Ma, G. Teng 2004. E-mail authorship mining based on SVM for computer forensic. In Proc. of the 3rd International Conference on Machine Learning and Cybernetics, Shanghai, China.

[48] V. Magri-Mourgues 2010. Distance intertextuelle et connexion lexicale : outils de catégorisation générique ou stylistique ? Approche expérimentale d'un corpus inédit : le corpus aragonien. JADT 2010.

[49] T. Mendenhall 1887. The characteristic curves of composition. Science, IX, pp : 237-249.

[50] F. Mosteller, D.L. Wallace 1964. Inference and disputed authorship: The Federalist. Addison-Wesley.

[51] J. Novak, P. Raghavan, A. Tomkins 2004. Anti-aliasing on the web. In Proc. of the 13th international conference on World Wide Web, pp : 30-39. ACM.

[52] F. Peng, D. Shuurmans, S. Wang 2004. Augmenting naive Bayes classifiers with statistical language models. Information Retrieval Journal, 7(1), pp : 317-345.

[53] J. Savoy 2012. Authorship attribution based on specific vocabulary. ACM Trans. Inf. Syst. 30, 2, May 2012.

[54] E. Stamatatos 2006. Authorship attribution based on feature set subspacing ensembles. International Journal on Artificial Intelligence Tools, 15(5), pp : 823-838.

[55] E. Stamatatos 2009. Intrinsic Plagiarism Detection Using Character n-gram Profiles. PAN'09, pp : 38-46.

[56] E. Stamatatos. 2009. A Survey of Modern Authorship Attribution Methods. JASIST.

[57] H. VanHaltern 2007. Author verification by linguistic profiling: An exploration of the parameter space. ACM Transactions on Speech and Language Processing.

[58] G. Yule 1938. On sentence-length as a statistical characteristic of style in prose, with application to two cases of disputed authorship. Biometrika, 30, pp : 363-390.

[59] G. Yule 1944. The statistical study of literary vocabulary. Cambridge University Press.

[60] Y. Zhao, J. Zobel 2005. Effective and scalable authorship attribution using function words. In Proceedings of the 2nd Asia Information Retrieval Symposium.

[61] G. Zipf 1932. Selected studies of the principle of relative frequency in language. Harvard University Press, Cambridge, MA.