Text-based Authorship Identification - A survey

Bushra Alhijawi*, Safaa Hriez†, Arafat Awajan‡
King Hussien School of Information Technology, Princess Sumaya University for Technology Amman, Jordan
Email: *bus20179001@std.psut.edu.jo, †killua92zoldyck@gmail.com, ‡awajan@psut.edu.jo

***Abstract*—The virtual world provides criminals with an anonymous environment to conduct malicious activities such as malware, sending ransom messages, spamming, theft intellectual property and sending ransom e-mails. All these activities are text in somehow. Therefore, there is a need for a tool in order to identify the author or creator of this illegal activity by analyzing the text. Text-based Authorship Identification techniques are used to identify the most possible author from a group of potential suspects of text. This paper is meant to explore the text-based authorship identification researches within the period 2007-2017. The researches were classified based on the application into email authorship, source code authorship, online text authorship, gender identification and online messages authorship. Also, the paper reviews and reports the datasets which used in the experiments of text-based authorship identification techniques. Finally, it reported the techniques which were used in authorship identification.**

***Index Terms*—Forensic Analysis, Authorship Analysis, Authorship Identification, Machine Learning, Datasets, Application, Writeprint, Features.**

I. Introduction

These days the researchers in the field of information security are interested in finding automated methods in order to determine the author of anonymous texts based on detecting some textual features. The presence of this field of research is a consequence of the information and communication technologies which provide an anonymous environment for users to conduct malicious activities in form of a textual communication [1] or plagiarism. The problem of anonymity in a text is addressed by applying Authorship Analysis **(AA)** techniques [1]. AA is the study of the linguistic and computational characteristics of the written documents of individuals [1]–[3]. It could be useful in many applications such as forensic analysis, online security, and making sure that the text is agreed with the organizational guidelines and style [4] (e.g. the text is a source code). The AA is studied from three main perspectives:

- Authorship Identification. Authorship Identification is applied to anonymous text to identify the most possible author of it from a group of authors [1], [5].
- Authorship Similarity Detection (**ASD**). ASD is used in order to determine if multiple texts are written by the same author, without knowing who the author is [1], [5].
- Authorship Characterization **(AC)**. AC is used to collect the demographic characteristics (e.g gender, age, educational level. etc.) of the potential author of an anonymous text [1], [5].

In that context, this survey aims to review the recent researches in Authorship Identification area during the last decade (2007 - 2017). In this study, the Authorship Identification researches are classified based on the application, dataset, year of publication and country. This survey reported the following:

- The most countries which are interested to work in text-based authorship Identification and in which year.
- The applications of the text-based authorship identification.
- The most datasets that are used to evaluate the authorship identification techniques.
- The most used techniques in Authorship Identification.

There are literature reviews on Authorship Identification (i.e. [3], [5]–[9]). However, these articles focus on earlier works or on a different perspective. For instance, Tamboli and Prasad, [3] focused on the Authorship Identification techniques that are published between 2001 and 2013 in term of features extraction. Nirkhi and Dharaskar [5] summarized Authorship Identification techniques used to identify authors of online messages during 2001-2012. In [6], the authors reviewed the papers in the period from 1998 to 2007. The works from 2000 to 2008 are reviewed by Koppel et al. [7] in term of three main scenarios: profiling problem, needle-in-a-haystack problem and verification problem. Chakraborty [8] provided an overview of the works in Authorship Identification of the documents written in Bengali. A compression between a set of distance measures that are used to reflect the stylistic similarity between authors and texts is done by Dinu and Popescu [9].

The rest of this paper is organized as follows. Firstly, Statistics of the selected researches are detailed. The features were proposed in the field of Author Identification which also known as Stylometric Features are presented in section III. The datasets which are used in the experiments are reviewed and discussed in section IV. Section V summarizes the techniques used in authorship identification. Next in section VI, the application of text-based Author Identification is presented. Finally, section VII conclude what was done in this research.

II. The Statistics of Researches based on Year, Country, Citations and Page Count

In order to identify relevant literature for this survey, a literature search was conducted on Google Scholar, ACM Digital Library, Springer Link, and ScienceDirect. Fifty papers from well-established and refereed journals and conferences were selected. As mentioned, The selected papers were published between 2007 - 2017 and most of them are between earlier

2011 - 2013 (i.e. 44%) and were published in journals (i.e. 44.52%). These research papers were sought using a combination of keywords as authorship, authorship identification, digital forensics, cybercrimes, email forensics, email misuse, authorship analysis, etc.

TABLE I
THE DISTRIBUTION OF RESEARCHES BASED ON YEAR

Year	Percentage of papers
2007	6%
2008	8%
2009	8%
2010	6%
2011	16%
2012	10%
2013	16%
2014	6%
2015	6%
2016	12%
2017	6%

In term of citation counts, 10% of the researches gained many citations (the maximum was 321 citations for [10]) and most researches had $1 - 19$ citations. Table II shows the distribution of researches based on citation count. The mean citation count was 40. From the reviewed researches, 4% had no citations. Citation counts were retrieved from Google Scholar at the end of 2017. Overall, the reviewed researches were full research papers. More than the half of researches (56%) had between eight and fifteen pages as shown in Table III. Another 34% had between four and seven pages. Most of the researches were done in the USA and Canadian universities with 30.61%. In the second place, India and Mexican universities did the same number of researches (6.45%). Table IV shows the distribution of researches based on country.

TABLE II
THE DISTRIBUTION OF RESEARCHES BASED ON CITATION COUNT

Citation count	Percentage of researches
0	4%
1-19	46%
20-39	18%
40-59	8%
60-79	8%
80-99	0%
100-119	6%
>120	10%

III. Stylometric Features

In order to predict the most appropriate author of an unknown text, the first step is to extract critical features that help in distinguish between authors. This section highlights some of these features.

Lexical Features

In the lexical features, the first and basic measures could be calculated for any text are the calculating of the words' length and the number of words in the text. These measures can be easily calculated for any language. [11] The *Vocabulary richness* is another lexical feature. This feature calculates the variation degree of the text's vocabulary. It is calculated by quantifying the number of the unique vocabularies, then divide this number over the total number of words in the whole text. [12] Another simple approach could be used to represent a text is to calculate the frequencies of the words (*word frequency*). The research of [13] showed that the most common words such as articles, pronouns and prepositions are the best features to distinguish between authors. *Word n-grams* have been proposed in order to take advantage of the combination of the words contextual information. It is the count of all possible combination of n words. For example, the term "get up", "he gets" and "get a job" are three occurrences of the word "get". [14] Another feature could be extracted from a text is the *errors* of that text. This feature extracts the words written with spelling errors. [15]

TABLE III
THE DISTRIBUTION OF RESEARCHES BASED ON PAGE COUNT

Page count	Percentage of researches
1-3	2%
4-7	34%
8-11	32%
12-15	24%
16-19	2%
20-23	0%
>24	6%

TABLE IV
THE DISTRIBUTION OF RESEARCHES BASED ON COUNTRY

Country	Percentage of papers
USA	16.1%
Canada	14.51%
Greece	4.84%
India	6.45%
Mexico	6.45%
Spain	4.84%
UK	3.25%
Denmark	3.25%
Romania	3.22%
Italy	3.22%
Others	33.87%

Character Features

According to this type of features, the text is considered as a sequence of characters. The measures could be calculated are the number of characters in the text, letter frequencies, upper and lower case characters, and punctuation marks and many others. [12] The most frequent *characters n-grams* approach could be applied with fixed or variable *n*. For example, the character 4-grams of the beginning of this paragraph would be |char|, |hara|, |arac|, |ract|, |acte|and so on. The advantage of this approach is that it is not affected by the errors of the text. For instance, the words simplistic and simplistic will output with a collection of common trigrams characters. Whereas these two terms have a different lexical presentation. [16]

Syntactic Features A more detailed method to represent the text is the syntactic features. The idea behind using these features is that the writers tend unconsciously to write in the same syntactic pattern. The extracting of these features is language dependent and need a particular parser for each natural language. [17] For example, the sentence *"Another try to use syntactic features was introduced by Stamatatos"* would be analyzed to: **NP**[*Another try*] **VP**[*to use*] **NP**[*syntactic features*] **VP**[*was introduced*] **PP**[*by Stamatatos*]. Where NP stands for noun phrase, VP stands for verb phrase and PP stands for a prepositional phrase. The measures which could be extracted are the count of the NP, VP and PP. Also, the length of NP, VP and PP, etc. [18] This could be done by labeling each phrase and then apply the measures on the output stream. One of the simple approaches is to use parts of speech tagger. This simple tool gives each word a tag depending on the contextual information. The researchers use the frequencies of the POS tag or n-grams frequencies of POS tag. The authors of [19] and [15] proposed an interesting syntactic feature based on the mistakes made by the writers including the mismatched tense, sentence fragments, etc.

Symantic Features

They are more detailed features; they consider the meaning of the text. [20] has proposed a tool in order to generate the semantic dependency graph. Another approach was proposed [21] to extract semantic measures. They found information about the hypernyms and synonyms of the phrases based on WordNet [22]. Furthermore, the researchers used latent semantic analysis to lexical features in order to find the similarities in the semantic between phrases. [23] Table V shows the basic stylometric features which are used in Authorship identification.

TABLE V
STYLOMETRIC FEATURES FOR AUTHORSHIP IDENTIFICATION

Features	Approaches
Lexical	Token-based (word length, sentence length, etc.)
	Vocabulary richness
	Word frequencies
	Word n-grams
	Errors
Character	Character types (letters, digits, etc.)
	Character n-grams (fixed length)
	Character n-grams (variable length)
Syntactic	Part-of-speech (POS)
	Errors
Semantic	Synonyms
	Semantic dependencies

IV. DATASETS

The datasets were used in the reviewed researches are classified into six main datasets. Table VI lists the reviewed researches by the dataset that is used in their works. The most used dataset is Enron E-mail Dataset. In 2001 Enron Company bankrupted because of the white collar fraud. Federal Energy Regulatory Commission made the e-mails of Enrons employees public. Enron dataset consists of more than 200,000 emails from about 150 employees. Two hundred words are the average number of words per message. They are written in the English language. The topics covered in these messages are ranging from personal chats to technical reports and business communications. [24] A total of 15 papers (30%) were used Enron dataset. They include [1], [10], [25]–[37].

Reuters Corpus Volume 1 (RCV1) is another dataset was used by [27], [38]–[41]. RCV1 consists of the following classes: ECAT (economics), MCAT (markets), GCAT (government/social), and CCAT (corporate/industrial). Each class consists of many subclasses. [42]

PAN competitions are conducted in 2012 using a dataset of English texts [43]. PAN authorship attribution development dataset was used in [44]–[48].

Blogger.com is a website where people could have profiles and can share anything with others [49]. Some researchers collected blog posts from this website and use them as a dataset to their works such as [50]–[52].

In order to check the authorship identification of source codes such as java codes. [53] used the dataset which is used by Lange and Mancoridis [54] and Bandara and Wijayarathna [55]. The dataset has java files written by ten authors and found in the Sourceforge website [56].

Some researchers used their own datasets by collecting texts from novels, books, twitter, articles, journals, poems, forums and other online sources. They include [4], [46], [57]–[75].

V. AUTHORSHIP IDENTIFICATION TECHNIQUES

After extracting the stylometric features of the text, the researchers used many techniques to identify the author of that text. The most used technique in the reviewed papers is Classification where 70% of the papers used classifiers in authorship identification process. There are many classifiers, the common used classifiers are Support Vector Machine (SVM) [10], [29], [31], [33], [35], [38], [39], [41], [57], [61], [63], [77], [78] , Naive Bayes [26], [52], [57] , Bayesian Network [26], [27], [32], [58] , Decision Tree (J48) [27], [33], [57] , Nearest Neighbors (k-NN) [39], [57], [60], [61] and Random Forest [64], [77]. The other technique used in authorship identification is Clustering where 12% of the reviewed researches used this technique. There are many clustering methods such as Expectation Maximization (EM) [34], k-means [1], [34], [50] and hierarchical agglomerative clustering [30]. Deep learning is another technique used in authorship identification. 4% of the papers used this technique; they are [40], [57], [70]. The remaining 14% of the papers used their own techniques [36], [44], [46], [59], [65], [70], [71]. Table VII shows the summary of the techniques used in Authorship Identification.

VI. THE APPLICATIONS OF TEXT-BASED AUTHORSHIP IDENTIFICATION

The text-based authorship identification is applied in many applications such as email authorship, source code, blog posts and discussion authorship. Table VIII shows the classification of the selected papers based on the application. It could be seen that most of the papers are proposed techniques to identify the

TABLE VI
LIST OF REVIEWED ARTICLES BY DATASETS

Dataset	Language	Papers
Enron E-mail Dataset [24]	English	[1], [10], [25]–[37]
Reuters Corpus Volume 1 [42]	English	[27], [38]–[41]
PAN authorship attribution development dataset [43]	English	[44]–[48]
NUS SMS Dataset [76]	English	[77]
Lange and Mancoridis [54] and Bandara and Wijayarathna [55]	Java	[53]
Blog posts	English	[50]–[52]
Their own	English, Arabic, Bengali, Java, Russian, others	[4], [46], [57]–[75]

TABLE VII
TECHNIQUES USED IN AUTHORSHIP IDENTIFICATION

Technique	Method	References
Classification	Support Vector Machine	[10], [29], [31], [33], [35], [38], [39], [41], [57], [61], [63], [77], [78]
	Naive Bayes	[26], [52], [57]
	Bayesian Network	[26], [27], [32], [58]
	Decision Tree (J48)	[27], [33], [57]
	Nearest Neighbors (k-NN)	[39], [57], [60], [61]
	Random Forest	[64], [77]
Clustering	Expectation Maximization (EM)	[34]
	k-means	[1], [34], [50]
	hierarchical agglomerative clustering	[30]
Deep Learning	Recurrent Neural Network	[40], [57], [70]
Others	Their own	[36], [44], [46], [59], [65], [70], [71]

TABLE VIII
THE APPLICATION OF TEXT-BASED AUTHORSHIP IDENTIFICATION

Application	Number of papers	Reference
Emails authorship	12	[10], [25], [26], [28], [29], [31], [32], [34]–[37], [47]
Source code authorship	5	[4], [10], [53], [65], [74]
Online text authorship	20	[29], [39]–[41], [45], [46], [50], [51], [57], [58], [60], [62], [64], [66]–[68], [71], [73], [75], [78]
Gender identification	2	[27], [33]
Instant messaging authorship	5	[30], [61], [70], [72], [77]
Others	5	[38], [44], [48], [52], [69]

author of an email or set of e-mails. The applications of the text-based authorship identification are summarized as follows:

- **Emails Authorship** [10], [25], [26], [28], [29], [31], [32], [34]–[37], [47].Emails are the most popular way to transmit information digitally with no authentication. Therefore, criminals use emails in abuse ways such as spam emails, phishing, email bombing, transmitting worms, forgery and email virus [37]. Email can be easily hacked and could be sent from a public internet cafe [37]. Therefore, the suitable solution in such cases is to examine the features of a malicious email to know its authorship from a list of suspects [37]. Most of the papers are published between 2011 and 2013 and they were done in the USA and Canadian universities. Many authors worked on this problem [10], [25], [26], [28], [29], [31], [32], [34]–[37], [47]. The common features between those techniques are that they ([10], [25], [26], [28], [29], [31], [32], [34]–[37], [47]) are classification-based techniques and used Enron email dataset in their experiments.
- **Source Code Authorship** [4], [10], [53], [65], [74]. The software may contain code from multiple authors due to the fact that the current software result of team efforts. Also, open source software written by multiple authors. Determining the authors of a piece of binary code from a set of known authors is the goal of source code authorship identification. But the question is why we need to identify the author of the code? The authorship identification is needed in this field for the following purposes:
  - The code's ownership is a suspect in cases such as in plagiarism or intellectual property infringement disputes [74].
  - Identifying the author/s of malware software. This is due to the fact that the malware is written by multiple

authors [79]. Malware creators share functional components by forming co-located teams [80] or through the Internet [81].

  – **Online text Authorship**. The web 2.0 technologies provide new opportunists to its users and facilitate publishing a large number of individually written electronic texts [50]. The need to identify the authors of those documents is becoming more important and challenging than before as may each user has various identities in the virtual world and maybe they behave differently in each context. Moreover, most other works in another field have focused on the case in which we need to identify the author of an anonymous document from a small set of candidate authors. But in this field, the set of known candidate authors is extremely large (i.e. may be many thousands) and might not even include the actual author [51]. Identifying the author of online text could be useful in various applications such as plagiarism, intellectual property and online security. Many authors were worked on this problem within the period 2010 - 2017 that are classified based on the text nature:
    * Blogs posts authorship [50], [51], [60].
    * Social network posts and comments [29], [57], [60].
    * Discussions authorship [64], [75].
    * General purpose [39]–[41], [44], [46], [58], [62], [66]–[68], [71], [73], [78].
- **Gender Identification** [27], [33]. The main goal is to determine if the author is a man or a woman.
- **Instant Messaging Authorship** [30], [61], [70], [72], [77]. Online messages provide users fast and easy communication way. Also, online messaging may be used for exchanging sensitive and secret information [30]. At the same time, online messaging can be misused by various means such as intimidation and an attacker may masquerade as a legitimate user. Therefore, in some cases there is a need to identify the author of the message.
- The text-based authorship identification can be used for many other purposes [38], [44], [48], [52], [69] such as handle class imbalance problem [38], detecting deviations in the writing style [44], literary works forensic [69] and authorship of translated text [48].

VII. Conclusion

This paper attempts to provide a survey of researches on authorship identification described in 2007 to 2017. The study gave statistics about the distribution of the researches based on the number of citations, the year and the country. It discussed the datasets and the features which were used in the reviewed researches. Also, it lists the applications where the authorship identification used in.

In conclusion, the features were used in the researches are a combination of four main features including lexical, Character, Syntactic and Semantic features. These features are extracted from datasets and from this study it has been observed that the most used dataset is Enron e-mail dataset where 30% of the papers used this dataset. The commonly used application for authorship identification is for emails and online texts. Five main applications of the text-based authorship identification were reported in this study: email authorship, source code authorship, online text authorship, gender identification and online messages authorship.

In term of the number of citation, most researches had $1-19$ citations with 46% of the researches and most of the researches were done in the USA and Canadian universities with 30.61% between the years 2011 and 2013.

The techniques used in authorship identification are classification, clustering and deep learning. The most used technique is classification-based where 70% of the selected papers used it.

References

[1] Farkhund Iqbal, Hamad Binsalleeh, Benjamin CM Fung, and Mourad Debbabi. A unified data mining solution for authorship analysis in anonymous textual communications. *Information Sciences*, 231:98–112, 2013.

[2] John F Burrows. Word-patterns and story-shapes: The statistical analysis of narrative style. *Literary & Linguistic Computing*, 2(2):61–70, 1987.

[3] Mubin Shaukat Tamboli and Rajesh S Prasad. Authorship analysis and identification techniques: A review. *International Journal of Computer Applications*, 77(16), 2013.

[4] Georgia Frantzeskou, Stephen G MacDonell, Efstathios Stamatatos, Stelios Georgiou, and Stefanos Gritzalis. The significance of user-defined identifiers in java source code authorship identification. *International Journal of Computer Systems Science and Engineering*, 2011.

[5] Smita Nirkhi and Rajiv V Dharaskar. Comparative study of authorship identification techniques for cyber forensics analysis. *arXiv preprint arXiv:1401.6118*, 2013.

[6] Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the Association for Information Science and Technology*, 60(3):538–556, 2009.

[7] Moshe Koppel, Jonathan Schler, and Shlomo Argamon. Computational methods in authorship attribution. *Journal of the Association for Information Science and Technology*, 60(1):9–26, 2009.

[8] Tanmoy Chakraborty. Authorship identification in bengali literature: a comparative analysis. *arXiv preprint arXiv:1208.6268*, 2012.

[9] Liviu P Dinu and Marius Popescu. Ordinal measures in authorship identification. In *Proc. SEPLN*, pages 62–66, 2009.

[10] Ahmed Abbasi and Hsinchun Chen. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)*, 26(2):7, 2008.

[11] Thomas Corwin Mendenhall. The characteristic curves of composition. *Science*, 9(214):237–249, 1887.

[12] Olivier De Vel, Alison Anderson, Malcolm Corney, and George Mohay. Mining e-mail content for author identification forensics. *ACM Sigmod Record*, 30(4):55–64, 2001.

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

[14] MC Rosa, VP Luis, MG Manuel, and R Paolo. Authorship attribution using word sequences. *Universidad Politécnica de Valencia*.

[15] Moshe Koppel and Jonathan Schler. Exploiting stylistic idiosyncrasies for authorship attribution. In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, volume 69, page 72, 2003.

[16] Tsukasa Matsuura and Yasumasa Kanada. Extraction of authors characteristics from japanese modern sentences via n-gram distribution. In *Discovery Science*, pages 315–319. Springer, 2000.

[17] Harald Baayen, Hans Van Halteren, and Fiona Tweedie. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3):121–132, 1996.

[18] Efstathios Stamatatos, Nikos Fakotakis, and George Kokkinakis. Automatic text categorization in terms of genre and author. *Computational linguistics*, 26(4):471–495, 2000.

[19] Joachim Diederich, Jörg Kindermann, Edda Leopold, and Gerhard Paass. Authorship attribution with support vector machines. *Applied intelligence*, 19(1):109–123, 2003.

[20] Michael Gamon. Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *Proceedings of the 20th international conference on Computational Linguistics*, page 611. Association for Computational Linguistics, 2004.

[21] Philip M McCarthy, Gwyneth A Lewis, David F Dufty, and Danielle S McNamara. Analyzing writing styles with coh-metrix. In *FLAIRS Conference*, pages 764–769, 2006.

[22] Adam Kilgarriff. Wordnet: An electronic lexical database, 2000.

[23] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391, 1990.

[24] Enron email dataset. Accessed: 2018-01-9.

[25] Farkhund Iqbal, Rachid Hadjidj, Benjamin CM Fung, and Mourad Debbabi. A novel approach of mining write-prints for authorship attribution in e-mail forensics. *digital investigation*, 5:S42–S51, 2008.

[26] Farkhund Iqbal, Hamad Binsalleeh, Benjamin CM Fung, and Mourad Debbabi. A unified data mining solution for authorship analysis in anonymous textual communications. *Information Sciences*, 231:98–112, 2013.

[27] Na Cheng, Rajarathnam Chandramouli, and KP Subbalakshmi. Author gender identification from text. *Digital Investigation*, 8(1):78–88, 2011.

[28] Marcelo Luiz Brocardo, Issa Traore, Sherif Saad, and Isaac Woungang. Authorship verification for short messages using stylometry. In *Computer, Information and Telecommunication Systems (CITS), 2013 International Conference on*, pages 1–6. IEEE, 2013.

[29] Marcelo Luiz Brocardo, Issa Traore, and Isaac Woungang. Authorship verification of e-mail and tweet messages applied for continuous authentication. *Journal of Computer and System Sciences*, 81(8):1429–1440, 2015.

[30] Smita Nirkhi, RV Dharaskar, and VM Thakare. Authorship verification of online messages for forensic investigation. *Procedia Computer Science*, 78:640–645, 2016.

[31] Sarwat Nizamani and Nasrullah Memon. Ceai: Ccm-based email authorship identification model. *Egyptian Informatics Journal*, 14(3):239–249, 2013.

[32] Michael R Schmid, Farkhund Iqbal, and Benjamin CM Fung. E-mail authorship attribution using customized associative classification. *Digital Investigation*, 14:S116–S126, 2015.

[33] Na Cheng, Xiaoling Chen, Rajarathnam Chandramouli, and KP Subbalakshmi. Gender identification from e-mails. In *Computational Intelligence and Data Mining, 2009. CIDM'09. IEEE Symposium on*, pages 154–158. IEEE, 2009.

[34] Farkhund Iqbal, Hamad Binsalleeh, Benjamin CM Fung, and Mourad Debbabi. Mining writeprints from anonymous e-mails for forensic investigation. *digital investigation*, 7(1):56–64, 2010.

[35] Rachid Hadjidj, Mourad Debbabi, Hakim Lounis, Farkhund Iqbal, Adam Szporer, and Djamel Benredjem. Towards an integrated e-mail forensic analysis framework. *digital investigation*, 5(3):124–137, 2009.

[36] A Pandian and AK Sadiq. Email authorship identification using radial basis function. *Int. J. Comput. Sci. Inform. Secu*, 9:68–75, 2011.

[37] Emad E Abdallah, Alaa E Abdallah, Mohammad Bsoul, Ahmed F Otoom, and Essam Al-Daoud. Simplified features for email authorship identification. *International Journal of Security and Networks*, 8(2):72–81, 2013.

[38] Efstathios Stamatatos. Author identification: Using text sampling to handle the class imbalance problem. *Information Processing & Management*, 44(2):790–799, 2008.

[39] Chunxia Zhang, Xindong Wu, Zhendong Niu, and Wei Ding. Authorship identification from unstructured texts. *Knowledge-Based Systems*, 66:99–111, 2014.

[40] LZ Wang. News authorship identification with deep learning, 2017.

[41] Adrián López-Monroy, Manuel Montes-y Gómez, Luis Villaseñor-Pineda, Jesús Carrasco-Ochoa, and José Martínez-Trinidad. A new document author representation for authorship attribution. *Pattern Recognition*, pages 283–292, 2012.

[42] David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397, 2004.

[43] Martin Potthast, Matthias Hagen, Tim Gollub, Martin Tippmann, Johannes Kiesel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. Overview of the 4th international competition on plagiarism detection. In *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[44] Gabriel Oberreuter and Juan D VeláSquez. Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style. *Expert Systems with Applications*, 40(9):3756–3763, 2013.

[45] Polina Diurdeva, Elena Mikhailova, and Dmitry Shalymov. Writer identification based on letter frequency distribution. In *Open Innovations Association (FRUCT), 2016 19th Conference of*, pages 24–30. IEEE, 2016.

[46] Helena Gómez-Adorno, Grigori Sidorov, David Pinto, and Ilia Markov. A graph based authorship identification approach. *Working Notes Papers of the CLEF*, 2015.

[47] Roman Kern, Christin Seifert, Mario Zechner, and Michael Granitzer. Vote/veto meta-classifier for authorship identification. In *CLEF 2011: Proceedings of the 2011 Conference on Multilingual and Multimodal Information Access Evaluation (Lab and Workshop Notebook Papers), Amsterdam, The Netherlands*, 2011.

[48] Steffen Hedegaard and Jakob Grue Simonsen. Lost in translation: Authorship attribution using frame semantics. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 65–70. Association for Computational Linguistics, 2011.

[49] Blog posts. Accessed: 2018-01-10.

[50] Haytham Mohtasseb and Amr Ahmed. Two-layer classification and distinguished representations of users and documents for grouping and authorship identification. In *Intelligent Computing and Intelligent Systems, 2009. ICIS 2009. IEEE International Conference on*, volume 1, pages 651–657. IEEE, 2009.

[51] Moshe Koppel, Jonathan Schler, and Shlomo Argamon. Authorship attribution in the wild. *Language Resources and Evaluation*, 45(1):83–94, 2011.

[52] Arvind Narayanan, Hristo Paskov, Neil Zhenqiang Gong, John Bethencourt, Emil Stefanov, Eui Chul Richard Shin, and Dawn Song. On the feasibility of internet-scale author identification. In *Security and Privacy (SP), 2012 IEEE Symposium on*, pages 300–314. IEEE, 2012.

[53] Upul Bandara and Gamini Wijayarathna. Source code author identification with unsupervised feature learning. *Pattern Recognition Letters*, 34(3):330–334, 2013.

[54] Robert Charles Lange and Spiros Mancoridis. Using code metric histograms and genetic algorithms to perform author identification for software forensics. In *Proceedings of the 9th annual conference on Genetic and evolutionary computation*, pages 2082–2089. ACM, 2007.

[55] Upul Bandara and Gamini Wijayarathna. A machine learning based tool for source code plagiarism detection. *International Journal of Machine Learning and Computing*, 1(4):337, 2011.

[56] Sourceforge. Accessed: 2018-01-10.

[57] Jenny S Li, Li-Chiou Chen, John V Monaco, Pranjal Singh, and Charles C Tappert. A comparison of classifiers and features for authorship authentication of social networking messages. *Concurrency and Computation: Practice and Experience*, 29(14), 2017.

[58] Richmond Hong Rui Tan and Flora S Tsai. Authorship identification for online text. In *Cyberworlds (CW), 2010 International Conference on*, pages 155–162. IEEE, 2010.

[59] Tanmoy Chakraborty and Prasenjit Choudhury. Authorship identification in bengali language: A graph based approach. In *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*, pages 443–446. IEEE, 2016.

[60] Jian Peng, Kim-Kwang Raymond Choo, and Helen Ashman. Bit-level n-gram based forensic authorship analysis on social media: Identifying individuals from linguistic profiles. *Journal of Network and Computer Applications*, 70:171–182, 2016.

[61] Tayfun Kucukyilmaz, B Barla Cambazoglu, Cevdet Aykanat, and Fazli Can. Chat mining: Predicting user and message attributes in computer-mediated communication. *Information Processing & Management*, 44(4):1448–1466, 2008.

[62] Alaa Saleh Altheneyan and Mohamed El Bachir Menai. Naïve bayes classifiers for authorship attribution of arabic texts. *Journal of King Saud University-Computer and Information Sciences*, 26(4):473–484, 2014.

[63] Ioannis Kanaris and Efstathios Stamatatos. Webpage genre identification using variable-length character n-grams. In *Tools with Artificial Intelligence, 2007. ICTAI 2007. 19th IEEE International Conference on*, volume 2, pages 3–10. IEEE, 2007.

[64] Mattia Samory and Enoch Peserico. Content attribution ignoring content. In *Proceedings of the 8th ACM Conference on Web Science*, pages 233–243. ACM, 2016.

[65] Xiaozhu Meng. Fine-grained binary code authorship identification. In *Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering*, pages 1097–1099. ACM, 2016.

[66] Jan Rygl and Aleš Horák. A framework for authorship identification in the internet environment. *Proceedings of RASLAN*, pages 117–124, 2011.

[67] Upendra Sapkota, Thamar Solorio, Manuel Montes-y Gómez, and Paolo Rosso. The use of orthogonal similarity relations in the prediction of authorship. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 463–475. Springer, 2013.

[68] Esteban Castillo, Ofelia Cervantes, Darnes Vilariño Ayala, David Pinto, and Saul León. Unsupervised method for the authorship identification task. *CLEF (Working Notes)*, 1180:1035–1041, 2014.

[69] Urszula Stańczyk and Krzysztof A Cyran. Machine learning approach to authorship attribution of literary texts. *International Journal of Applied Mathematics and Informatics*, 1(4):151–158, 2007.

[70] Marco Cristani, Giorgio Roffo, Cristina Segalin, Loris Bazzani, Alessandro Vinciarelli, and Vittorio Murino. Conversationally-inspired stylometric features for authorship attribution in instant messaging. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 1121–1124. ACM, 2012.

[71] Nuno Homem and Joao Paulo Carvalho. Authorship identification and author fuzzy fingerprints. In *Fuzzy Information Processing Society (NAFIPS), 2011 Annual Meeting of the North American*, pages 1–6. IEEE, 2011.

[72] Maarten Lambers and Cor Veenman. Forensic authorship attribution using compression distances to prototypes. *Computational Forensics*, pages 13–24, 2009.

[73] Marius Popescu and Cristian Grozea. Kernel methods and string kernels for authorship analysis. In *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[74] Jay Kothari, Maxim Shevertalov, Edward Stehle, and Spiros Mancoridis. A probabilistic approach to source code authorship identification. In *Information Technology, 2007. ITNG'07. Fourth International Conference on*, pages 243–248. IEEE, 2007.

[75] Sindhu Raghavan, Adriana Kovashka, and Raymond Mooney. Authorship attribution using probabilistic context-free grammars. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 38–42. Association for Computational Linguistics, 2010.

[76] Tao Chen and Min-Yen Kan. Creating a live, public short message service corpus: the nus sms corpus. *Language Resources and Evaluation*, 47(2):299–335, 2013.

[77] Esther Villar-Rodriguez, Javier Del Ser, Miren Nekane Bilbao, and Sancho Salcedo-Sanz. A feature selection method for author identification in interactive communications based on supervised learning and language typicality. *Engineering Applications of Artificial Intelligence*, 56:175–184, 2016.

[78] Stefan Ruseti and Traian Rebedea. Authorship identification using a reduced set of linguistic features. In *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[79] Fernando De La Cuadra. The geneology of malware. *Network Security*, 2007(4):17–20, 2007.

[80] Mandiant Intelligence Center. Mandiant 2013 threat report. *WEB-2013-MNDT-RPT-M-Trends-2013 LP.html, 2013. Mandiant White Paper*, 2013.

[81] Ahmed Abbasi, Weifeng Li, Victor Benjamin, Shiyu Hu, and Hsinchun Chen. Descriptive analytics: Examining expert hackers in web forums. In *Intelligence and Security Informatics Conference (JISIC), 2014 IEEE Joint*, pages 56–63. IEEE, 2014.