Recent Approaches on Authorship Attribution Techniques – An Overview

Siddharth Swain*, Gaurav Mishra and C. Sindhu
Department of Computer Science and Engineering,
SRM University, Kattankulathur - 603 203
Kancheepuram, Tamil Nadu, India.
Email: siddharthswain1001@gmail.com

***Abstract*— Authorship analysis deals with the identification of authors which is a problem of text data mining and classification. There are numerous techniques and algorithms that have been published so far, in the field of stylometry. In this regard, the primary objective of the present review is to provide the status of the different studies carried out on authorship analysis based on the important research contributions. The authors have mainly focused on each of the article selected for review (2010-16), by summarizing the authorship detection fields, the corpus, the features of authorship analysis, and authorship attribution techniques used which could provide a platform to distinguish between different research contributions and the diverse techniques used in classifying the author's texts. The details on common tools and authorship attribution techniques published in the recent past would be of importance to the concerned researchers as an aid towards text data mining and future growth in the study of authorship attribution.**

Keywords—*stylometry, writing style, authorship analysis, authorship attribution, authorship detection, the corpus, features of authorship analysis, authorship attribution techniques, plagiarism, natural language processing, text-mining*

I. Introduction

Authorship analysis is the process of examining the characteristic features of a piece of text in order to conclude its authorship which has its roots in linguistic research called stylometry [1]. The individual authors have distinctive style or ways of writing as well as speaking, and there exists a long history of linguistic and stylistic investigations into authorship attribution [2]. Therefore, authorship analysis and its attribution have various societal applications and considered as the potential areas of research in the field of stylometry which deal with identification of the author(s) of any given text by using several different approaches or techniques of text data mining. Obviously, longer the text, always better is the identification accuracy expected. The subject area of stylometry or authorship recognition/detection has several related fields of research such as: (i) Authorship Attribution (AA) or identification, i.e. identifying the author(s) of a set of different texts; (ii) Authorship Characterization (AC), i.e. detection of sociolinguistic attributes like gender, age, occupation and educational level of the author; (iii) Authorship Verification (AV), i.e. checking whether a text or set of similar texts is written or not written by an author; (iv) Authorship Discrimination (AD), i.e. checking of two different texts whether those are written by the same author; (v) Plagiarism Detection (PD), i.e. search for the sentence(s) or paragraph(s) that are reproduced from the text(s) of one or more than one author; (vi) Text Indexing and Segmentation (TIS), i.e. when several texts are issued from different authors, which are concatenated in a form of a global book or forum text and (vii) Authorship De-Identification (ADI), i.e. to identify features which can properly capture an author's writing style. Some of the commonly used stylometric features to capture an author's writing style are namely, (i) lexical features, i.e. text is viewed as a sequence of tokens grouped in to sentences; (ii) syntactic features, i.e. patterns used to form sentences; (iii) structural features, i.e. how an author organizes the structure of the document; (iv) content-specific features, i.e. characterization of certain activities, discussion forum or interest groups with few key words and so on.

The AA is a classical classification problem [3] of interest in stylometry. In general, AA can be clearly defined as the task of inferring characteristics of a document's author from the textual characteristics of the document itself [4]. Further it is emphasized that, such an analysis can always be performed either on a piece of handwritten text or on a digital document. In this digital era when there is a rapid growth in the worldwide network(s) and the power of cloud computing, AA becomes handy and essential with newer and faster techniques. Hence, the primary objective of this paper is to provide an overview of the different studies done on authorship analysis and AA techniques. We have attempted to review the selected papers that are recently published (2010-2016) by appropriately tabulating the different analysis, approaches or techniques (more than 40 different methods) put forward by various researchers in the field of AA without considering any specific literature or language alone. It is well established that, if a text is suspected to be an amalgam of different sources and by different authors, determination of AA becomes a complex task, whereas such a document can be of sufficient commercial or legal interest [5]. In addition, the generic or the common tools and AA techniques published so far in the recent past and which would be of importance in the

* Student Member of IEEE Society .

978-1-5090-5686-6/17/$31.00 ©2017 IEEE

concerned field of research for the future growth of AA have been discussed and highlighted in this paper.

II. METHODOLOGY

Authorship analysis dealing with the identification of authors is a problem of data mining and classification [6]. There are numerous methods and algorithms that have been published to understand the subject better. Keeping this in mind, a total of 46 selected papers including three review articles are chosen by us for review, and they have been accepted as the proven and well established database of publications on the present study. In an earlier review, already referred above [7] on authorship analysis, the individual papers (1992-2004) were tabulated chronologically for the authorship analysis which has been adopted here as such (Table-I). In addition to this table another table (Table-II) has been prepared carefully for indicating different AA techniques which are reported in different studies/papers. As far as possible, the abbreviations of most techniques as shown in Table-II would be used here subsequently. It is intended that such studies may further suggest or generate an urge for newer and futuristic techniques, and may trigger further work.

As explained above, table-I presents the articles reviewed in this study chronologically keeping the important details such as: the authorship detection field, the corpus, respective domain, the feature types, language and the AA techniques or methods used by the authors of each paper considered. A variety of domains such as publications, art, literature, rhymes, prose, narratives, conferences, news articles, religion, politics, history, law, health, sports, acting, romance, film, economy, social, travel & tourism, trade, business, goods & services, science, technology, bio-medical, forensic, computer, command history (IT), SMS messages, SPAM emails, tweets, chat logs, web forum posts, and general text/miscellaneous/ random topics distributed uniformly are covered. The major feature types used in the review are namely lexical features (F1) i.e. character and word based features, syntactic features (F2), structural features (F3) and, content-specific features (F4). This survey includes more than twelve different languages namely Arabic, Bengali, Chinese, English, German, Japanese, Latin, Persian, Portuguese, Russian, Thai and Turkish including inter-language, multi-language texts. The corpus details and different AA techniques used are also mentioned in table-I.

III. RESULTS AND DISCUSSION

Initial works on authorship analysis goes back to 19th century with the study of Shakespeare's plays followed by the statistical studies during the first half of 20th century [1]. The study of the "Federalist Papers" was considered the most influential work in AA. For the past 50 years, linguists, computer scientists, and scholars of the humanities have been jointly developing automated and efficient methods for AA and analysis based on the style of writings [8]. Before authorship analysis measures the textual features, it avoids distinguishing between texts written by different authors. The literature clearly emphasize that, all authors possess specific and unique peculiarities of their habit that influence the form and content of their writing style.

Authorship analysis has been used successfully to analyze the provenance of source code files in several studies [9]. It has potential to be used to link malware with other malware written by the same authors, helping investigations, classification, deterrence and detection. Electronic mails too have increasingly replaced all written modes of communications for important correspondences including personal and business transactions [10]. Most of the time, an e-mail is given equal significance as that of a signed document. Hence, email impersonation through compromised accounts has become a major threat. Electronic text stylometry aims at identifying the writing style of authors of electronic texts such as electronic documents, blog posts, tweets, etc. Identifying such styles is quite attractive for identifying authors of disputed e-text, identifying their profile attributes or even enhancing services such as search engines and recommender systems [11]. Similarly, SMS messaging is also a popular media of communication. Because of its popularity and privacy, it could be used for many illegal purposes [12]. The Internet is a decentralized structure that offers speedy communication, and has a global reach and provides anonymity, a characteristic invaluable for committing illegal activities [13]. Authorship analysis has been used fruitfully in solving all such issues as described below under four sub-headings (Table-I) including the various techniques used for AA. Each of the paper under discussion is addressed separately based on the importance of its contributions by segregating them under separate heads as follows:

A. Authorship Detection Field

As explained above, the various authorship detection fields such as **AA**, **AC**, **AV**, **AD**, **PD**, **TIS** and **ADI** of authorship detection [1] are given in Table-I against each reference. It may be noted that some studies deal with more than one to all most all fields of authorship detection which associates text to authors based on their style of writing. Over the past two decades, it has been extensively studied by researchers for natural language processing [14].

- The paper [15] brings the notion of AA on the astroturfing problem by collecting quantities of data from social media sites and analyzing the putative individual authors to see if they appear to be the same person. Here, the analysis comprises of a binary **CnG** which was earlier shown to be effective for accurately identifying authors.
- AA of the ancient texts [16] written by ten Arabic travelers using the Arabic dataset namely AAAT which contains 3 short texts from every book showed good AA performance with an optimal score of 90%. Moreover, this investigation has also revealed interesting results concerning the Arabic language.

TABLE-I: THE DETAILS OF SELECTED PAPERS (2010-2016) ON AUTHORSHIP ANALYSIS

AA: Authorship Attribution (Identification), **AC:** Authorship Characterization, **AV:** Authorship Verification, **AD:** Authorship Discrimination, **PD:** Plagiarism Detection, **TIS:** Text Indexing and Segmentation, **ADI:** Authorship De-Identification; **F1:** Lexical features, **F2:** Syntactic features, **F3:** Structural features, **F4:** Content-specific features; **CnG:** Common n- Gram, **SVM:** Support Vector Machines, **NBC**: Naive Bayesian Classifier, **POS:** Part of Speech Tagging, **MLP:** Multi Layer Perceptron or other Neural Networks (**NN**), **DF (**Distance Functions like Manhattan, Cosine, Euclidean, Stamatatos & Camberra distances), **C 5/4.5** or **(DT)** Decision Trees, **SMO:** Sequential Minimal Optimization, **k-NN:** k Nearest Neighbour Algorithm, **Tf-idf:** Term Frequency- Inverse Document Frequency Method, **BOW:** Bag of Words Model, **SCAP** Source Code Author Profile Method, **MC:** Markov Chains Method, **RF:** Random Forests Method, **CS:** Cosine Similarity Method, **NUANCE** Method, **RLP:** Recentred Local Profile Method, **Graph Based** Model/Technique, **IGN:** Information Gain Method, **WAN:** Word Adjacency Network Model, **k- Means** Clustering Method, **Chi-Square** Method, **KLD** (*Kullback–Leibler Divergence Method*), **MLE** (Maximum Likelihood Estimation).

Table Index: Column (2), Ref – References; Column (3), ADF - Authorship Detection Field

Year	Ref	ADF	Corpus	Domain	Features	Language	Techniques
2010	[19]	AA	6 datasets, No: of authors (5-100), Avg. Posts per author (17-162)	Web Forum Posts	F1, F2, F4, POS, Perplexity values from CnG	English	SVM, NN, BN, NB, DT
2011	[4]	AA	3000 short articles, 30 authors, Taken from 15 Brazilian Newspapers, Avg. article-600 tokens 350 Hapax (words occurring once)	Law, Sports, Literature Politics, Tourism, Health, etc.	F1,F2	Portuguese	SVM, Multi-objective Genetic Algorithm, BOW
2012	[2]	AA	10 groups belonging to 10 ancient authors, Each group has 3 texts by a same author (Avg. text length: 550 words)	Travel	F1, F4, CnG, Rare Words	Arabic	SVM, SMO
2012	[5]	TIS	Pair of biblical books, Set of blog posts, 4 Columnists writing for the New York Times	Identically distributed, Miscellaneous Topics	F1, F2	English (Language Independent)	SVM
2012	[45]	PD	Thirteen authors, 126 documents containing an average of 4933 words per author and 500 words per document	Miscellaneous Topics, Topic-Independent	F1, F2, F3, CnG	English (Language Translated)	NB, SVM, SMO
2012	[45]	PD	Thirteen authors, 126 documents containing an average of 4933 words per author, and 500 words per document	Miscellaneous Topics, Topic-Independent	F1, F2, F3, CnG	English (Language Translated)	NB, SVM, SMO
2012	[46]	AA	Three Novels each written by the writers, Hermann Hesse, and Stephan Zweig	---	F1, F2, F3	German	Committee Machine, MLP, k-NN
2012	[29]	TIS	40 novels & proses for each of the 5 famous authors, shortest: 943 & longest: 143712 characters.	Miscellaneous Topics	F1, F2, CnG	Chinese	Ward's Method, k-Means, DF, KLD
2012	[23]	AA	IRC messages from the Ubuntu IRC channel between 2004 &2012, Twitter & Web Forums	Miscellaneous to Technical Topics	F4, CnG, Rare Words	English	Inverse Author Frequency, SCAP, RLP
2012	[36]	AA, AC	Chat Logs; 341 posts spread over 17 days from April to September 2006 from a US based chat-room.	Miscellaneous Topics, Forensic Analysis (Perspective)	F1, F2, F3, F4, CnG	English	Chi-Square, SVM, NB, MC, Bayesian Regression
2013	[44]	AA, AD, AC	Works written by Mark Twain, Herman Melville, Shakespeare, Austen, Allen, etc.	Literary	F1, F2	English	SVM, WAN
2013	[22]	AA	Streaming Data, IMDB62 database, there are 62 authors with a thousand of comments for each of the authors	Movie-Related such as Reviews	F1, F2, POS, Tag CnG, Tf-idf	English	SVM, BOW
2013	[21]	AA, AV, PD	C/C++/Java programs from Planet Source Code Website, Student-submitted solutions from RMIT University	Source Codes (Programs/ Software)	---	Programming Languages: C, C++, Java	Burrow's Method, SCAP
2013	[18]	AA	Works of famous & contemporary Iranian writers (each document: min 800 - max 1000 words)	Romance, Literary, Social, Satire, etc.	F1, F2	Persian	SVM, k-NN, DT, BOW, Modified Tf-idf

2013	[35]	AA	Short historical texts that are written by ten ancient Arabic travellers, This Arabic dataset, which were collected by the authors in 2011 called AAAT dataset	Travel, History, Narratives, Conferences, etc.	F1,F2, CnG, Rare Words	Arabic	DF, MLP, SMO, SVM & Linear Regression.
2013	[34]	AA, AV	Russian texts were taken from proceedings (2006 to 2012) of the Dialogue international conference on Computational Linguistics	Technical, General, etc.	F1, F2, F3, CnG, POS	English, Russian	SVM, Confusion Matrix
2013	[20]	AA, TIS	Four different datasets of IRC Logs were used, “perverted justice”, “krijin”, “irclogs”, and “omegle”	Chat Logs	F4, Stop Words	English	*KLD, MLE*
2013	[16]	AA	10 groups belonging to 10 ancient authors, each group has 3 texts by same author (Avg. text length: 550 words)	Conferences Travel, History, etc.	F1, CnG	Arabic	MLP, DF, SVM, SMO
2013	[33]	AA	Entire non obfuscation portion of the extended Brennan-Greenstadt	Miscellaneous Topics	F1, F2, F3, POS, CnG, Write-prints Feature Set	Latin, English Inter-Language	Own Algorithm
2013	[12]	AA	NUS - one of the largest SMS message corpus, has more than 50 thousand messages from multiple cultures in Asia	SMS Messages	Unigram Method (CnG)	Multi-Language	CS, DF
2013	[13]	AA	Real world data sets compiled from a global system of spam traps were used, consisting of 27 clusters, varying in size from a maximum of 505 emails to only one email	SPAM Emails	F1, F2, F3, CnG	Multi-Language	NUANCE, k-Means
2013	[43]	AA	2000 messages taken from an SMS corpus with each author capped at 50 messages	Text and Instant Messages	F1, F2, F3	Language-Independent	Discrete NB, Gaussian and a Custom Classifier
2014	[42]	AA	Online messages from pantip.com web-board & two other fan pages	Acting, Politics, etc.	F1, F2, F3	Thai	SVM, DT
2014	[26]	AA, AV, PD	7231 programs from open source, samples from textbooks	Source Codes (Programs/ Software)	CnG	Programming Languages: C++, Java	SCAP
2014	[27]	AA	2,000 tweets per user from over 8 million Japanese users by using Twitter API during Jan 2013 to Dec 2013	Tweets, Short Messages	POS-Tag-combined CnG	Japanese	CS, Biased weighting technique for CnG
2014	[17]	AA	Command Line History collected over 25 years. Eg: Greenberg & Schonlau datasets	Command History	CnG, Distinct First words, Semantic	Unix Language	NB, DT, Maximum Entropy
2014	[37]	AA, AV, AC	2484 articles of the NIPS conferences from 1987 to 2003	Technical, Scientific, General, etc.	F1, F2, F3	English	MC, RF, Own algorithm
2014	[32]	ADI	180 & 169 abstracts collected from 5 authors each in Computer Science & Bio-medical field respectively from Microsoft Academic Research Website	General, Computer Science, Biomedical Fields, etc.	F1, F2, F3, F4	English	MLP, SVM, k-NN, RF
2014	[9]	AA, AV, TIS, PD	Zeus Botnet Source Code has been used	Source Codes (Programs/ Software)	CnG, File Purpose, Explicit Markers	Programming Languages: C++, C, etc.	RLP, CS, NUANCE
2015	[41]	AA	CCAT C10 document collections created by Efstathios Stamatatos from the *Reuters Corpus Volume 1(*RCV1) collection	Miscellaneous Sources: News articles, Publications	F1, F2, F3, POS	English	Graph Based, SVM
2015	[11]	AA	PAN 12 database has been used which provides various types of Stylometry problems	General	F1, F2, F3, CnG, POS	---	---
2015	[10]	AA	Enron public email dataset; belonging to 158 users	Personal, Business, etc.	F1, F2, CnG	English	One class SVM, Graph Based, Probability Model, Inclusive Compound Probability Model

2015	[31]	AA	Literary works by 33 poets from different eras has been used.	Rhymed, Measured, Prose, etc.	F1, First Words, Rhymes	Arabic	MC, Chi Square, IGN
2015	[30]	AA, AC	Data from multilingual marketplace forum- Black Market Reloaded, contains 92,333 posts from 8,348 users posted in 12,923 threads	Illegal Trade of goods and services, related areas, etc	F1, F2, F3, Character CnG, Time based	Multi-Language	CS, SVM
2015	[25]	AA	53,205 Tweets from Twitter API by 20 Arabic users from Middle East	Art, Religion, Politics, etc.	---	Arabic	BOW, Tf-idf, NB
2016	[28]	AA, AV	2 short text databases and 2 long text databases taken from newspapers, Gutenberg project, etc.	Literary, General, etc.	F2	Portuguese, English	SVM, Own Algorithm
2016	[39]	AA	Two corpora WMPR-AA2016-A and WMPR-AA2016-B of Persian poem having 12 candidate authors	Literary	Uni, Bi, and tri-gram weighting+ Inverse Document Frequency	Persian	Modified Language Modelling
2016	[3]	AA	168 English Texts of famous authors from Project Gutenberg	Literary Works		English	Normalized Relative Compression
2016	[14]	AV	2836 novels from 136 different authors from Project Gutenberg	Literary Works	Standard, Modified & Partial Hausdorff distance	English	DF, Probabilistic k-NN & SVM, Baseline- LSH based pruning
2016	[38]	AA	40 Novels written by 8 authors between 1835 and 1922 from Project Gutenberg	Literary Works	F2, Frequency of Motifs	English	WAN, Motifs, SVM, k-NN, NB, DT
2016	[6]	AA	22,000 Turkish newspaper articles which belong to different genres between 1995 and 2015	Life, Economy, Political, etc.	F1, F2, F3,F4	Turkish	Artificial NN, Levenberg Marguardt based classifier
2016	[24]	AA	Extended-Brennan-Greenstadt corpus which has samples of 45 unique authors, collected from Amazon's Mechanical Turk.	General, Literary, etc.	F1, F2, Semantic & Frame based, POS	English	Frame Semantic Method, SVM, Bag of Frames Method
2016	[15]	AA, AV TIS, PD	Comments in news and opinion sites like ABC Drum, Guardian online, etc.	General, Miscellaneous Topics	CnG	English	Binary N-gram Analysis
2016	[7]	AA	Bengali Blog corpus of 3000 passages written by three authors	Miscellaneous Random Topics	F1, Stop Words, CnG, Term Frequency	Bengali	NB, SMO, DT, MLP, Tf-idf, IGN

- Using standard algorithms and feature sets inspired by natural language, AA demonstrates that individual users can be identified with a high degree of accuracy through their command-line behavior [17]. For a 50 user configuration, the study found feature sets that can successfully identify users with more than 90% accuracy.
- Paper [18] is on evaluation of the effects of textual features on AA accuracy. In this paper, several classification algorithms were used on corpora with 2, 5, to 20 and 40 different authors and a comparison was performed. The evaluation results showed that the information about the used words and verbs were most reliable criteria for AA tasks and also NLP based features were more reliable than BOW based features.
- Paper [19] demonstrated a two stage approach for combining unsupervised and supervised learning approaches for performing AA on web forum posts. During the first stage, the approach focused on using clustering techniques to make an effort to group the data sets into stylistically similar clusters. The second stage involved using the resulting clusters from stage one as features to train different machine learning classifiers.
- The authors of article [20] worked on AA for a novel set of documents, namely online chats. Although the problem of AA has been extensively investigated for different

document types, from books to letters and from emails to blog posts, it was the first approach on conversational documents using statistical models. The investigators experimentally demonstrated the unsuitability of the classical statistical models for conversational documents and proposed a novel approach which could achieve a high accuracy rate up to 95% for hundreds of authors.

- Article [21] compared the two most effective methods such as Burrows and SCAP and it includes an extension of the study. The original comparative study only considered anonymized data while the replicated study considered both anonymized and non-anonymized data. The original comparative study indicated that the Burrows Method outperformed all other methods – including the SCAP method – by a considerable margin.
- In another study on AA on streaming data [22], the concept of novel authors occurring in streaming data source such as evolving social media was addressed. The study focused on what happens if new authors are added into the system by time? Moreover, the study also focused on problems that some of the authors may not stay, and may disappear by time or may reappear after a while.
- The study of AA for IRC messages using inverse-author-frequency [23] revealed that this application was difficult due to the short messages and repeated information. To improve the accuracy the authors had applied higher weights to features used by fewer authors for the first time.
- Authorship detection of SMS messages using unigrams [12] was a proven method which specifically tried to compare how well the algorithms worked under less amount of testing data and a large number of candidate authors against controlled tests with less number of authors and selected SMSes with large number of words.
- In paper [24], authors present a technique that incorporates the use of semantic frames as a method for authorship attribution. They hypothesized that it provides a deeper view into the semantic level of texts, which is an influencing factor in a writer's style. They used a variety of online resources in a pipeline fashion to extract information about frames within the text.

B. *The Corpus*

The details of text data sets used by each of the publication have been tabulated in Table-I along with its domain of research in the next column. The corpus details and domain of each paper need no further explanation. However, the specific contribution of each article is worth mentioning.

- Paper [6] introduced a new authorship identification process based on Artificial Neural Network (ANN) model using embedded stylistics features. It is well known that stylistics features mostly depend on the topic or genre of the articles. Here the basic dataset contains 22,000 Turkish newspaper articles which belong to different genres. The results indicated that 97% success rate was achieved with Levenberg Marguardt based classifier. The study suggested that the corpus presented in this work for the first time might contribute towards not only authorship identification but also other identification purposes.
- The investigators of article [14] proposed a novel solution by modeling AA as a set similarity problem to overcome some existing limitations. They conducted experiment extensively on a real datasets collected from an online book archive. Their results showed that in comparison to existing stylometry studies, their solution on natural language processing could handle a larger number of documents of different lengths written by a larger pool of candidate authors with very high accuracy.
- Several state-of-the-art features have been tested for the Arabic language and particularly for very short texts. In one of the rare works, two particularities such as Arabic language and small text size were emphasized [2]. In this paper, the authors had investigated the task of AA on very old Arabic texts which were written by ten ancient Arabic travellers. Several features such as characters n-grams and word n-grams (**CnG)** were used as input of a **SMO-SVM**. The experiments of AA on this database showed interesting results with a classification precision of 80%.
- The paper [25] was on using big data analytics for authorship authentication of Arabic tweets. None of the previous works of Arabic text had addressed the unique challenges associated with tweets and large-scale datasets. The results showed that the testing accuracy was not very high (61.6%) as expected.
- Source code AA [26] is the task of determining the author of source code whose authorship is not explicitly known. One specific method of source code for AA that was proven to be extremely effective is the SCAP method. In this article, several alternate approaches were found to perform better than the baseline approach used in the SCAP method. The approach that performed the best was empirically shown to improve the performance from 91% to 97.2% measured as a percentage of documents correctly attributed using a data set consisting of 7,231 programs written in Java and C++.
- The internet security issues require authorship identification for all kinds of internet contents. However, authorship identification for microblog users is much harder than other documents because microblog texts are too short [27]. Moreover, when the number of candidates become large, i.e., big data, it will take longer time for identification. This article solved such problems and the experimental results showed that, this method was used successfully to identify the authorship with 53.2% of precision out of 10,000 microblog users in almost half the execution time of previous methods.

C. *Features of authorship analysis*

The most important features (Table-I) considered by various researchers Lexical features (**F1**), Syntactic features (**F2**), Structural features (**F3**), Content-specific features (**F3**) and the associated languages studied are accordingly included for classification and analysis against each publication which are highlighted as below:

- The paper [28] presented syntactic features for the AA to literary texts in verification and identification. The authors of this article used a classification method based on dissimilarity that has been successfully applied in cases of AA. To evaluate the writer-dependent model and writer-independent model, they tested both approaches based on polytomy and dichotomy. They also evaluated impact of the redundancy by varying the number of references for short and long texts. They achieved significant results higher than 90% in both languages (Table-II) for verification and above 75% for identification.
- Paper on text clustering on authorship attribution based on the features of punctuations usage [29] proposed a method of extracting writing characteristics of various authors based on their usage of punctuation marks. Here, comparative analysis has been done between the text clustering effects of the proposed method and character Bigram method using 200 articles of five well-known modern writers.
- AA on dark marketplace forums, i.e. paper [30] has proposed classification set-ups for two tasks related to user identification namely alias classification and AA. The authors of this paper revealed that for both tasks, they achieved high accuracy results using a combination of character-level n-grams, stylometric features and timestamp features of the user posts.
- Paper [31] presented an Arabic poetry as an AA task. Several features such as characters, sentence length, word length, rhyme, and first word in sentence were used as input data for Markov Chain methods. The data was filtered by removing the punctuation and alphanumeric marks that were present in the original text. Subsequently, a set of thirty-three poets from different eras was used which produced interesting results with classification precision of 96.96%.
- Extraction of significant features that represent an author's style from the available concise emails is a big challenge in email AA [10]. Hence, it was proposed to use a graph-based model to precisely extract the unique feature set of the authors using one-class SVM classifier to deal with the single class sample data that consists of only true positive samples. Two classification models were successfully designed in this study and they were compared well.
- The main challenge of authorship de-identification is to identify features which can properly capture an author's writing style. The investigators of another paper [32] had chosen as a combination of all four (**F1 to F4**) features to represent author's style. The study concluded that among four well-known classifiers, MLP achieved the best performance for authorship de-identification.
- In the article on "Identifying Authors with Style" [33], the researchers analyzed a feature set previously introduced using a tool and corpus already available. Decomposing the set, they identified the features that seem to have contributed the most accurate performance.
- Paper [34] is on the measurement of certain expressible features of writing style, and its uses include the characterization of authors for recognition in cases of text whose authorship is disputed or unknown. This work builds upon previous investigations into the success of a particular feature set on a particular corpus.
- AA of short historical Arabic texts based on Lexical Features [35] could investigate the authorship of several short historical texts that are written by ten ancient Arabic travelers.
- Another study [36] could successfully analyze the application of different AA techniques to chat log from a forensic perspective.

D. *Authorship Attribution Techniques*

By representing large corpora with concise and meaningful elements, topic-based generative models aim to reduce the dimension and understand the content of documents [37]. Those techniques originally analyzed on words in the documents, but their extensions currently accommodate meta-data such as authorship information, which has been proved useful for textual modeling. In this review of 43 papers, the two AA techniques such as **CnG** and **SVM** are used/studied by the authors of 23 and 21 publications respectively (Table-II). Minimum six to eight studies used one of the techniques such as **NBC, POS, MLP/ANN, DF, C-5/4.5** and **DT.** The methods such as **SMO**, **k-NN**, **Tf-idf**, **BOW**, **SCAP**, **MC**, **RF** and **CS** are used or studied by the authors of three to five publications. Maximum of two papers used one of the seven other techniques namely **NUANCE**, **RLP**, **Graph Based**, **IGN**, **WAN**, **k-Means** and **Chi-Square**. Each of the other techniques listed in the bottom most row of Table-II are used only in one publication each, as shown. Brief details on the AA techniques are as follows:

- Paper [3] illustrated the performance of a compression-based measure that relies on the notion of relative compression, besides comparing with recent approaches that use multiple discriminant analysis and support vector machines. The authors of this paper attained 100% correct classification of the data sets used showing consistency between the compression ratio and the classification performance.
- The goal of paper [38] was to apply the concept of motifs, i.e. recurrent interconnection patterns, in the AA task. The absolute frequencies of all thirteen directed motifs with three nodes were extracted from the co-occurrence networks and used as classification features. The effectiveness of these features was verified with four machine learning methods. The authors of this paper have found that function words play an important role in these recurrent patterns.
- Paper [39] evaluated the experimental results by four approaches such as unigram, bigram, trigram; and

modified language modeling by using two Persian poem corpora as WMPR-AA2016-A Dataset and WMPRAA2016-B dataset. Results showed that modified language modeling revealed better performance than other approaches.

- The authors of paper [11] presented an evaluation of Random Forests (RF) on the problem domain of AA. They have taken advantage of RF's robustness against noisy features by extracting a diverse set of features from evaluated e-texts. Interestingly, the resultant model achieved the highest classification accuracy in all problems, except one occasion where it misclassified only a single instance.
- Paper [40] revealed that, since function words are independent of content, their use tends to be specific to an author and that the relational data captured by function WANs is a good summary of stylometric fingerprints. Attribution accuracy is observed to exceed the one achieved by methods that rely on word frequencies alone.
- Paper [41] used graphs for representing document sentences for solving the problem of AA. The experiments which are presented here could attain approximately 79% accuracy indicating that the graph-based representation could be a way of encapsulating various levels of natural language descriptions.
- The SAT model [37] outperformed the AT model for identifying authors of documents written by either single authors or multiple authors with a better Receiver Operating Characteristic (ROC) curve and a significantly higher Area Under Curve (AUC).
- Paper [42] presents a framework to identify the authors of Thai online messages. The identification is based on 53 writing attributes and the selected algorithms are support vector machine (SVM) and C4.5 decision tree. Experimental results indicated that the overall accuracies achieved by the SVM and the C4.5 were 79% and 75%, respectively.
- ChatSafe [43] is an author attribution system intended for use with short message based communication, i.e. instant messaging or SMS. The authors have presented a modified Bayesian classifier, used internally by ChatSafe, which improves on the accuracy of a standard Bayesian classifier.
- The AA using function words adjacency networks [44] has presented a method based on relational data between function words. These are content independent words that help in defining grammatical relationships.
- In paper [45], the authors have investigated the effects of machine translation tools on translated texts and the accuracy of authorship and translator attribution of translated texts. The study revealed that, the more translation performed on a text by a specific machine translation tool, the more effects unique to that translator, are observed. This study achieved 91% to 91.5% accuracy.
- The paper on AA using committee machines with k-nearest neighbors rated voting [46] presents that the determination of the author of a text can become an extraordinarily complex and sensitive job due to its relatively difficult feature extraction phase and highly nonlinear nature. Hence, this paper proposed a classification tool using committee machines consisting of MLP neural networks to identify the author of a text.

TABLE-II STUDIES DEALING WITH DIFFERENT AUTHORSHIP ATTRIBUTION TECHNIQUES

SN	***Techniques/Methods: References***	***Total***
1	**CnG (**Common n-Gram): [2], [7], [9], [10], [11], [12], [13], [15], [16], [17], [19], [22], [23], [26], [27], [29], [30], [33], [34], [35], [36], [39], [45]	**23**
2	**SVM** (Support Vector Machines): [2], [3], [4], [5], [10], [16], [18], [19], [22], [24], [28], [30], [32], [34], [35], [36], [38], [41], [42], [44], [45]	**21**
3	**NB (**Naive Bayes Classifier): [7], [17], [19], [25], [36], [38], [43], [45]	**8**
4	**POS (**Part of Speech Tagging): [11], [19], [22], [24], [27], [33], [34], [41]	**8**
5	**DF (**Distance Functions like Manhattan, Cosine, Euclidean, Stamatatos & Camberra distances): [9], [12], [14], [16], [27], [29], [30], [35]	**8**
6	**MLP (**Multi-Layer Perceptron or other Neural Networks): [6], [7], [16], [19], [32], [35], [46]	**7**
7	**C-5/4.5** or Decision Trees **(DT)**: [7], [17], [18], [19], [38], [42]	**6**
8	**SMO** (Sequential Minimal Optimization): [2], [7], [16], [35], [45]	**5**
9	**k-NN** (k Nearest Neighbour Algorithm): [14], [18], [32], [38], [46]	**5**
10	**Tf-idf** (Term Frequency- Inverse Document Frequency Method): [7], [18], [22], [25]	**4**
11	**BOW (**Bag of Words Model): [4], [18], [22], [25]	**4**
12	**SCAP** (Source Code Author Profile Method): [21], [23], [26]	**3**
13	**MC** (Markov Chains Method): [31], [36], [37]	**3**
14	**RF** (Random Forests Method): [11], [32], [37]	**3**
15	**CS** (Cosine Similarity Method): [12], [27], [30]	**3**
16	**NUANCE** Method: [9], [13]	**2**
17	**RLP** (Recentred Local Profile Method): [9], [23]	**2**
18	**Graph Based** Model/Technique: [10], [41]	**2**
19	**IGN** (Information Gain Method): [31], [38]	**2**
20	**WAN** (Word Adjacency Network Model): [38], [44]	**2**
21	**k-Means** Clustering Method: [13], [29]	**2**
22	**Chi-Square** Method: [31], [36]	**2**
23	**KLD** (Kullback–Leibler Divergence Method): [20], [29]	**2**
Other Techniques: Normalized Relative Compression [3], Multi-objective Genetic Algorithm [4], Levenberg Marguardt based Classifier [6], Inclusive Compound Probability Model [10], Baseline-LSH based pruning [14], Probability Model [10], Maximum Entropy Method [17], Bayesian Network [19], **MLE** (Maximum Likelihood Estimation) [20], Burrow's Method [21], Frame Semantic/ Bag of Frames Method [24], Ward's Method [29], Confusion Matrix [34], Linear Regression [35], Bayesian Regression [36], Motifs [38], Modified Language Modelling [39], Gaussian & Custom Classifier [43], Committee Machines [46]		

IV. CONCLUSIONS

The primary objective of this review is to provide the status of the different studies carried out on authorship analysis based on the important research contributions. Accordingly the authors have presented an overview of the recent approaches on authorship analysis and attribution techniques based on a total of 46 selected publications including three review articles. They have attempted for an analysis based on, (i) authorship detection fields, (ii) the corpus, (iii) features of authorship analysis, and (iv) authorship attribution techniques by preparing one single table (Table-I) for all the articles considered. The table could help in dealeaneating various kind of contributions in different articles and the diverse techniques used to classify the author's texts which are discussed briefly based on each of the article included in this study. The review covers several different text domains and languages. Hopefully, the details presented on various tools and authorship attribution techniques published in the recent past would help researchers as an aid towards text data mining and to work towards future growth of the subject. It appears from the contributions presented in different papers that there are numer of techniques and algorithms that have been published so far and they can provide promising results in the field of stylometry, especially on authorship attribution.

The second table could provide first hand information on studies dealing with different authorship attribution techniques. The most fequently used authorship attribution techniques were Common n-Gram and Support Vector Machines followed by Naive Bayesian Classifier, Part of Speech Tagging, Multi Layer Perceptron or other Artificial Neural Networks, Distance Functions like Manhattan, C-5/4.5 and Decision Trees. Currently, the Internet and Mobile network is a decentralized structure that surely offers speedy communication, but provides ample scope for committing illegal activities. In this regard, authorship analysis being a problem of text data mining and classification, Artificial Neural Networks appears to be most promising technique for future applications.

Acknowledgments

The authors convey their sincere thanks to Dr. B. Amutha, Professor and Head, Department of Computer Science and Engineering, SRM University, Kattankulathur and Dr. Revathi Venkataraman, Professor of the same department for their encouragement and timely support provided. The work carried out in this paper is a part of 8th Semester project work (B. Tech. in Computer Science and Engineering) of the first two authors under the guidance and supervision of Professor C. Sindhu.

References

[1] Sara El Manar El Bouanani and Ismail Kassou. Article: Authorship Analysis Studies: A Survey. *International Journal of Computer Applications* 86(12):22-29, January 2014.

[2] S. Ouamour and H. Sayoud, "Authorship attribution of ancient texts written by ten arabic travelers using a SMO-SVM classifier," *2012 International Conference on Communications and Information Technology (ICCIT)*, Hammamet, 2012, pp. 44-47. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6285841&isnumber=6285764

[3] A. J. Pinho, D. Pratas and P. J. S. G. Ferreira, "Authorship Attribution Using Relative Compression," *2016 Data Compression Conference (DCC)*, Snowbird, UT, 2016, pp. 329-338. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7786177&isnumber=7786136

[4] P. Varela, E. Justino and L. S. Oliveira, "Selecting syntactic attributes for authorship attribution," *The 2011 International Joint Conference on Neural Networks*, San Jose, CA, 2011, pp. 167-172. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6033217&isnumber=6033131

[5] N. Akiva and M. Koppel, "Identifying Distinct Components of a Multi-author Document," *2012 European Intelligence and Security Informatics Conference*, Odense, 2012, pp. 205-209. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6298832&isnumber=6298809

[6] O. Yavanoglu, "Intelligent authorship identification with using Turkish newspapers metadata," *2016 IEEE International Conference on Big Data (Big Data)*, Washington, DC, 2016, pp. 1895-1900. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7840809&isnumber=7840573

[7] S. Phani, S. Lahiri and A. Biswas, "A machine learning approach for authorship attribution for Bengali blogs," *2016 International Conference on Asian Language Processing (IALP)*, Tainan, Taiwan, 2016, pp. 271-274. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=& arnumber=7875984&isnumber=7875919

[8] A. Rocha *et al.*, "Authorship Attribution for Social Media Forensics," in *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 1, pp. 5-33, Jan. 2017. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=& arnumber=7555393&isnumber=7726079

[9] R. Layton and A. Azab, "Authorship Analysis of the Zeus Botnet Source Code," *2014 Fifth Cybercrime and Trustworthy Computing Conference*, Auckland, 2014, pp. 38-43. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7087326&isnumber=7087313

[10] Novino Nirmal. A, Kyung-Ah Sohn and T. S. Chung, "A graph model based author attribution technique for single-class e-mail classification," *2015 IEEE/ACIS 14th International Conference on Computer and Information Science (ICIS)*, Las Vegas, NV, 2015, pp. 191-196. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp? tp=&arnumber=7166592&isnumber=7166553

[11] M. Khonji, Y. Iraqi and A. Jones, "An evaluation of authorship attribution using random forests," *2015 International Conference on Information and Communication Technology Research (ICTRC)*, Abu Dhabi, 2015, pp. 68-71. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=& arnumber=7156423&isnumber=7156393

[12] R. Ragel, P. Herath and U. Senanayake, "Authorship detection of SMS messages using unigrams," *2013 IEEE 8th International Conference on Industrial and Information Systems*, Peradeniya, 2013, pp. 387-392. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6732015&isnumber=6731935

[13] M. Alazab, R. Layton, R. Broadhurst and B. Bouhours, "Malicious Spam Emails Developments and Authorship Attribution," *2013 Fourth Cybercrime and Trustworthy Computing Workshop*, Sydney NSW, 2013, pp. 58-68. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6754642&isnumber=6754625

[14] S. Nutanong, C. Yu, R. Sarwar, P. Xu and D. Chow, "A Scalable Framework for Stylometric Analysis Query Processing," *2016 IEEE 16th International Conference on Data Mining (ICDM)*, Barcelona, 2016, pp. 1125-1130. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7837960&isnumber=7837813

[15] J. Peng, R. K. K. Choo and H. Ashman, "Astroturfing Detection in Social Media: Using Binary n-Gram Analysis for Authorship Attribution," *2016 IEEE Trustcom/BigDataSE/ISPA*, Tianjin, 2016, pp. 121-128. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp. jsp?tp=& arnumber=7846937&isnumber=7846883

[16] S. Ouamour and H. Sayoud, "Authorship attribution of ancient texts written by ten Arabic travelers using character N-Grams," *2013 International Conference on Computer, Information and Telecommunication Systems (CITS)*, Athens, 2013, pp. 1-5. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6705713&isnumber=6705707

[17] F. Khosmood, P. L. Nico and J. Woolery, "User identification through command history analysis," *2014 IEEE Symposium on Computational Intelligence in Cyber Security (CICS)*, Orlando, FL, 2014, pp. 1-7. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7013363&isnumber=7013356

[18] R. Ramezani, N. Sheydaei and M. Kahani, "Evaluating the effects of textual features on authorship attribution accuracy," *ICCKE 2013*, Mashhad, 2013, pp. 108-113. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6682828&isnumber=6682797

[19] S. R. Pillay and T. Solorio, "Authorship attribution of web forum posts," *2010 eCrime Researchers Summit*, Dallas, TX, 2010, pp. 1-7. [Online]. available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5706693&isnumber=5706677

[20] G. Inches, M. Harvey and F. Crestani, "Finding Participants in a Chat: Authorship Attribution for Conversational Documents," *2013 International Conference on Social Computing*, Alexandria, VA, 2013, pp. 272-279. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=& arnumber=1556355&isnumber=33093

[21] M. F. Tennyson, "A Replicated Comparative Study of Source Code Authorship Attribution," *2013 3rd International Workshop on Replication in Empirical Software Engineering Research*, Baltimore, MD, 2013, pp. 76-83. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=& arnumber=6664734&isnumber=6664717

[22] S. E. Seker, K. Al-Naami and L. Khan, "Author attribution on streaming data," *2013 IEEE 14th International Conference on Information Reuse & Integration (IRI)*, San Francisco, CA, 2013, pp. 497-503. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6642511&isnumber=6642428

[23] R. Layton, S. McCombie and P. Watters, "Authorship Attribution of IRC Messages Using Inverse Author Frequency," *2012 Third Cybercrime and Trustworthy Computing Workshop*, Ballarat, VIC, 2012, pp. 7-13. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6498422&isnumber=6498412

[24] R. Hinh, S. Shin and J. Taylor, "Using frame semantics in authorship attribution," *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Budapest, 2016, pp. 004093-004098. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7844873&isnumber=7844217

[25] J. Albadarneh *et al*., "Using Big Data Analytics for Authorship Authentication of Arabic Tweets," *2015 IEEE/ACM 8th International Conference on Utility and Cloud Computing (UCC)*, Limassol, 2015, pp. 448-452. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=& arnumber=7431455&isnumber=7431374

[26] M. F. Tennyson and F. J. Mitropoulos, "Choosing a profile length in the SCAP method of source code authorship attribution," *IEEE SOUTHEASTCON 2014*, Lexington, KY, 2014, pp. 1-6. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6950705&isnumber=6950640

[27] S. Okuno, H. Asai and H. Yamana, "A challenge of authorship identification for ten-thousand-scale microblog users," *2014 IEEE International Conference on Big Data (Big Data)*, Washington, DC, 2014, pp. 52-54. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7004491&isnumber=7004197

[28] P. Varela, E. Justino, A. Britto and F. Bortolozzi, "A computational approach for authorship attribution of literary texts using sintatic features," *2016 International Joint Conference on Neural Networks (IJCNN)*, Vancouver, BC, 2016, pp. 4835-4842. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7727835&isnumber=7726591

[29] M. Jin and M. Jiang, "Text clustering on authorship attribution based on the features of punctuations usage," *2012 IEEE 11th International Conference on Signal Processing*, Beijing, 2012, pp. 2175-2178. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnnumber=6492012&isnumber=6491878

[30] M. Spitters, F. Klaver, G. Koot and M. v. Staalduinen, "Authorship Analysis on Dark Marketplace Forums," *2015 European Intelligence and Security Informatics Conference*, Manchester, 2015, pp. 1-8. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7379716&isnumber=7379706

[31] A. F. Ahmed, R. Mohamed, B. Mostafa and A. S. Mohammed, "Authorship attribution in Arabic poetry," *2015 10th International Conference on Intelligent Systems: Theories and Applications (SITA)*, Rabat, 2015, pp. 1-6. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7358411&isnumber=7358370

[32] J. Hurtado, N. Taweewitchakreeya and X. Zhu, "Who wrote this paper? Learning for authorship de-identification using stylometric featuress," *Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014)*, Redwood City, CA, 2014, pp. 859-862. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7051981&isnumber=7051718

[33] L. M. Stuart, S. Tazhibayeva, A. R. Wagoner and J. M. Taylor, "On Identifying Authors with Style," *2013 IEEE International Conference on Systems, Man, and Cybernetics*, Manchester, 2013, pp. 3048-3053. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6722273&isnumber=6721750

[34] L. M. Stuart, S. Tazhibayeva, A. R. Wagoner and J. M. Taylor, "Style Features for Authors in Two Languages," *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, Atlanta, GA, 2013, pp. 459-464. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6690051&isnumber=6689975

[35] S. Ouamour and H. Sayoud, "Authorship Attribution of Short Historical Arabic Texts Based on Lexical Features," *2013 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, Beijing, 2013, pp. 144-147. [Online]. Available: http://ieeexpl ore. ieee.org/stamp/stamp.jsp?tp=&arnumber=6685672&isnumber=6685639

[36] F. Amuchi, A. Al-Nemrat, M. Alazab and R. Layton, "Identifying Cyber Predators through Forensic Authorship Analysis of Chat Logs," *2012 Third Cybercrime and Trustworthy Computing Workshop*, Ballarat, VIC, 2012, pp. 28-37. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6498425&isnumber=6498412

[37] N. Pratanwanich and P. Lio, "Who Wrote This? Textual Modeling with Authorship Attribution in Big Data," *2014 IEEE International Conference on Data Mining Workshop*, Shenzhen, 2014, pp. 645-652. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7022657&isnumber=7022545

[38] V. Q. Marinho, G. Hirst and D. R. Amancio, "Authorship Attribution via Network Motifs Identification," *2016 5th Brazilian Conference on Intelligent Systems (BRACIS)*, Recife, 2016, pp. 355-360. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7839612&isnumber=7839541

[39] S. Vazirian and M. Zahedi, "A modified language modeling method for authorship attribution," *2016 Eighth International Conference on Information and Knowledge Technology (IKT)*, Hamedan, 2016, pp. 32-37. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7777783&isnumber=7777746

[40] S. Segarra, M. Eisen and A. Ribeiro, "Authorship Attribution Through Function Word Adjacency Networks," in *IEEE Transactions on Signal Processing*, vol. 63, no. 20, pp. 5464-5478, Oct.15, 2015. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7140830&isnumber=7243391

[41] E. Castillo, D. Vilariño, O. Cervantes and D. Pinto, "Author attribution using a graph based representation," *2015 International Conference on Electronics, Communications and Computers (CONIELECOMP)*, Cholula, 2015, pp. 135-142. [Online]. Available: http://ieeexplore .ieee.org/stamp/stamp.jsp?tp=& arnumber=7086940&isnumber=7086811

[42] R. Marukatat, R. Somkiadcharoen, R. Nalintasnai and T. Aramboonpong, "Authorship Attribution Analysis of Thai Online Messages," *2014 International Conference on Information Science & Applications (ICISA)*, Seoul, 2014, pp. 1-4. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6847369&isnumber=6847317

[43] J. A. Donais, R. A. Frost, S. M. Peelar and R. A. Roddy, "Summary: A System for the Automated Author Attribution of Text and Instant Messages," *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, Niagara Falls, ON, 2013, pp. 1484-1485. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6785915&isnumber=6785655

[44] S. Segarra, M. Eisen and A. Ribeiro, "Authorship attribution using function words adjacency networks," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, 2013, pp. 5563-5567. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6638728&isnumber=6637585

[45] A. Caliskan and R. Greenstadt, "Translate Once, Translate Twice, Translate Thrice and Attribute: Identifying Authors and Machine Translation Tools in Translated Text," *2012 IEEE Sixth International Conference on Semantic Computing*, Palermo, 2012, pp. 121-125. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6337093&isnumber=6337075

[46] A. O. Kusakci, "Authorship attribution using committee machines with k-nearest neighbors rated voting," *11th Symposium on Neural Network Applications in Electrical Engineering*, Belgrade, 2012, pp. 161-166. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6419997&isnumber=6419933