ISSN (e): 2250 – 3005 || Volume, 08 || Issue, 9|| Sepetember – 2018 ||
International Journal of Computational Engineering Research (IJCER)

A Survey On Authorship Attribution Approaches

Sreenivas Mekala[1], Vishnu Vardan Bulusu[2], Raghunadha Reddy T[3]
[1] Research Scholar, Rayalaseema University, Kurnool, Asst. prof, Dept. of IT, SNIST, Hyderabad
2 professor, Dept of CSE, Vice-Principal, JNTUHCEM.
[3]Associate Professor, Dept of IT, Vardhaman College of Engineering, Hyderabad
Corresponding Author: Sreenivas Mekala

ABSTRACT: The goal of Authorship Attribution (AA) is to take decisions about an author of unique chunk of text. AA is a variant of classification problem and it differs from classical text classification problem. It is a text analysis process that has different techniques namely, Authorship Identification, Authorship Profiling, Authorship verification, Authorship clustering, Authorship Diarization, and Plagiarism Detection. In this study a brief survey of Authorship attribution technique are presented. AA is a process, given a document and a set of candidate authors, determine who among them wrote the document. Plenty of accessible electronic writings are available in different areas on web and sometimes the authors of the text are unknown. The primary goal of the survey is to address various stylomeric features used and Attribution techniques based on the text corpus written by different authors.
KEYWORDS: Authorship Identification, Authorship Profiling, Authorship verification, Authorship clustering, Plagiarism Detection, text classification and stylomeric features.

Date of Submission: 01-10-2018 Date of acceptance: 13-10-2018---

---

I.
II. INTRODUCTION

The primary goal of authorship attribution is that when two documents are given; find out whether the two documents are written by the two different authors or single author [1]. Authorship Attribution (AA) can be treated as a classification problem. Text Categorization is labeling of documents to a set of predefined classes. AA is a process of recognizing the authorship of a given document, over a corpus whose authorship is known [2]. The need of the study is to trace out if any known hostile person giving warnings to government bodies, verifying the authenticity of suicide, writing harassing messages, copyright disputes. Based on these many researchers proposed and addressed many techniques to find out the unknown authors. Researchers proposed variety of features to identify the writing style characteristics of authors [3]. AA becomes an important problem due to fast growing electronic text throughout world. It is helpful when two or more individual authors declared to have written something or when no one is ready to agree who is actual author. AA can also called as author identification [4], closed class problem [5] states that from the given a set of authors, exact author of the text is definitely from one among those authors set, Open-class problem [5] states that from the given a candidate set of authors, exact author of the text may be outside of the, categorization problem, vanilla authorship attribution in [6], needle-in-a-haystack problem in [7]. Precisely authorship attribution is defined as one author is assigned to a piece of text of unknown authorship, given a set of authors for whose text samples are available. In Data Mining, this can be treated as a multi-class single-label text categorization problem [8].This problem is also known as authorship identification. Exclusive studies are made on authorship attribution by Stamatatos et.al.[9].Different tasks of authorship analysis are listed below.

1. Author verification; determine whether a selected piece of text content was written by a specific author or not.
2. Plagiarism detection, measuring the closeness between two texts.
3. Author profiling or characterization, obtain information concern to sex, age, education etc. of the author of a given text.

Apart from the traditional application to literary research[10], [11] , AA is applied to many different diverse areas like intelligence[12], criminal law[13] , civil law[14] , computer forensics[15].

III. RELATED WORK

The basic idea of authorship attribution supported by Machine learning techniques and statistical methods is that identify some features from the given text and by using those features one can identify text written by different authors. In late 1880's, initial efforts were made on authorship attribution on the plays of Shakespeare. Many

researchers Yule, G.U , Zipf et.al. contributed in the middle of 20th century towards authorship attribution. The history of Authorship attribution based on statistical methods is much older. Mosteller et.al. contributed considerably on the attribution of 'The Federalist Papers'. The statistical language modelling techniques of Mosteller et.al. was the first work published in this area. In contrast to human expert-based systems this study has initiated the computer based and computer assisted Authorship attribution techniques. E.Stamatatos proposed that the authorship attribution techniques and are divided into 2 groups. The method and feature used are mainly divided into two sets. Apart from the different generative and discriminative methods, some of the features like Lexical, syntactic and semantic features are also used. From then the study on authorship attribution was purely concentrating on identifying the features to quantify the writing styles of different authors. This type of research called as 'stylometry'. Mosteller et.al. and Holmes et.el. proposed different measures some of them are vocabulary richness functions, character frequencies, word frequencies, sentence length, word length. Ganascia et.al proved that the performance of frequencies of function words better than the sequential rules. Words whose purpose is to contribute to the syntax rather than the meaning of a sentence are called as function words some example are prepositions, pronouns, articles. Feature which will commonly used are called shallow features such as character features, N-gram features, tokens. Some features like rewrite rule frequencies, part-of-speech require some deep analysis.

A. **Stylomeric Features**

The simplest way of representing the text is continuous set of words, these words are called tokens. Sentence is continuous set of words. A token may be character, string, special character, literal, digit or a number. The simplest measures of text in authorship attribution are count of words and sentences and their counts.

Stylometry works on the assumption that every author has specific style of writing and it has some specific feature. These features provide a ground to identify the author .In general the features of stylometry are as follows. Count of sentences in a text, count of words of in a document, average count of words in a given text, average word length, count of periods, count of exclamation marks, count of commas, count of colons, count of semicolons.

There are numerous tools to identify the different features. Lexical features, token-based features are identified by a Tokenizer with sentence splitter, Vocabulary richness feature is identified by Tokenizer with a Stemmer and Lemmatizer, Word frequencies is calculated with the help of Tokenizer, Word n-grams is identified with the help of tolls like Tokenizer, Orthographic Spell Checker.

Character features like Character type digits, Letters are identified by mapping to Character dictionary, Character n-grams (fixed length) is identified by Feature selector, Character n-grams (variable length) are identified by Text compression tools.

Syntactic features like Part-Of-Speech are identified by Sentence Splitter, Tokenizer, Pos Tagger. Sentences and Phrase Structures are identified by chunker, Rewrite rules frequencies are identified by Text Chunker Pos-Tagger, Tokenizer, and Sentence Splitter. Errors are identified by Pos-Tagger, Syntactic Spell Checker, Text Chunker, Sentence Splitter, Tokenizer, Full parser, Partial parser.

Semantic features like Synonyms identified by Thesaurus, Pos-Tagger, and Tokenizer. Semantic dependencies are measured by Pos-tagger, Text chunker, Tokenizer, Sentence Splitter. Functional features are identified by Specialized Dictionaries, Sentence Splitter, Semantic Parser, Pos-Tagger, Partial parser, Tokenizer.

Application-Specific features like structural features are identified by HTML parser, Specialized Parsers. Content-specific features are identified by Lemmatizer, Stemmer, and Tokenizer. Language-Specific features are identified by Lemmatizer, Stemmer, and Tokenizer.

The process of authorship attribution involves in Pre-Processing, Feature extraction, Selection of features, Model design and performance measurement with the help of different metrics. There are mainly 3 different types of features. They are Syntactic features, Lexical features and Structural features.

B. **Lexical Features**

In general word or character based features are considered as Lexical features. Some of the Word-Based Lexical features are count of all words, count of words in a Sentence, length of the word and Vocabulary Richness, these metrics contains number of words which appears only one time called as hapax legomena and appears two times is called as hapax dislegomena. Different type of lexical features are special characters, letter frequency, content words, misspellings, character n-grams as in [12], [8], sentence length in [16] **,** Verbal Phrases, phrase length in [17], function words in [18] , words per phrase type, phrase types [9], function word-token ratios, type-token ratio, character n-grams in [19], unigrams , word n-grams in [20], words bigrams or sequences **,** Function word frequencies, POS trigrams or sequences of 3 in [17**],** Pos- Bigrams in [21], Pos-Trigrams in [17], Pos-Tags in [8], PCFG-obtained POS in [22], Complexity measures with Pos in [21], Function words in [12], non-function words in [22], 1024-character sequences in [18], syntactically classified punctuation in [13]. The Structural features font size, font colour as in [12], word length distribution and vocabulary richness in [8] ,

word distribution, punctuation distribution, punctuation frequency in [20], syntactically classified punctuation , syntactic structure in [13],Punctuation marks, special use words in [16], spelling errors, word form errors, most frequent types in [10], emotions, frequency of lemmas, frequency of negative words , hyperlinks, font formatting in [23].Word-based features include statistical metrics such as hapax legomena and hapax dislegomena, average length of a word, average length of sentence, type token-ratio, number of bi-gram, tri-gram, quad-gram characters, and Vocabulary rich number measure such as Sichel S , Honore R, Yule K, Simpson D, Entropy measures are considered for the attribution as in [24].

C. **Character Based Features**

Generally text is viewed as sequence of characters. Some of the character-based features are number of letters, uppercase characters, digits, white spaces, special characters. Character-based lexical features consisting of aggregate of all characters, number of characters in each sentence, number of characters in each word and no. of e occurrences of each letter. Syntax is defined as the structure used in the construction of sentence. This type of features consisting of the rules used to form sentences like function words, punctuation. Usage pattern of function word is a useful feature for authorship identification. In this way, different character level measures were characterized, it includes digit count, alphabet count, count of lowercase and uppercase characters, frequency of letter , count of punctuation marks [25], [26], [4]. Ian Baker et.al. in [27] used the uppercase letters to all character , white spaces to all character, tab spaces to all characters, Upper case letters to small letters, proportion of numeric information is also used as features in the text. The character level n-gram features are important in dealing with the character based features. The most repeatedly occurring character n-grams will play major role in stylistic purposes. Numerous variety of tool are not required to attain most repeated n-grams, and attainment process is fully independent of language used.However Stamatatos et.al. addressed that when compare to word-based approach, degree of representation is substantially raised. The reason is very clear that n-grams will catch up unessential information and no. of character n-grams are require to symbolize a unique lengthy word (e.g., |and_|, |_and|) . Magdalena et.al. in [28] considered frequency of the most common 4-grams character. In the work of Erwan Moreau [29] Character unigrams, trigrams and 5-grams for text characterization are considered. Julio Villena is used n-gram based character sequences, based on distance among histograms for each attribute. Octavia-Maria found that the best tf-idf features were observed at character-level where n-gram ranges from 2 to 6 and after this threshold. Compression-based approaches by Khmelev et.al Z. in [30] and Marton et al. are considered as special context of using character information. The principal thought is compression model produces one text to compress another text. Vocabulary Diversity is measuring the richness or diversity of an author's vocabulary is also used as a discriminating feature. In information Retrieval, Bag of Words is all the words, excluding the stop words are used in document vector. Some of the function words conjunction, pronoun are utilized as a segregating features of authors. Neural networks based automated pattern recognition is also used [12], [13], it not much used because it includes training a neural network which is useful to identify the authors style. Neural Network are useful in learning the text style and useful in text categorization. This technique is useful in authorship attribution, Neural Networks will learn and classify one author from reaming set of authors. This type of techniques involves in finding the average word length in the form of no for characters and letters [13], syllables and avg. no. of words in sentence [15]. These measures are proved to be not sufficient some measures like the no. of words appearing with given frequency in a text and type-token ratio are used.

D. **Syntactic Features**

A function word is a word which is significantly less meaningful content. These are considered as structured grammatical words in English which has a structural relationship with other words in a sentence. These function words includes the grammatical aspects of English such as pronouns (she, they), determiners (the, that), prepositions (in, of), auxiliary verbs (be, have), modals (may, could), conjunctions (and, but) and quantifiers (some, both). Some researchers used function words as features and proved that the male authors use more prepositions when compared to females. Gilad Gressel extracted around seven features from the text which includes the grammatical aspects such as adjectives, nouns, determiners, pronouns, adverbs and foreign words. The morpho syntactic information tags were assigned to every word token based on the contextual information. This is a process carried out by a Part of speech (Pos) Tagger. This Pos Tagger identifying the styles of the authors quite accurately by using POS tag n-gram frequencies or POS tag frequencies [29, 31, 32] from the unrestricted text. POS tag information provides the structural analysis of sentences and never reveals the fact about the combination of words to form phrases or high level structures. While identifying the demographic features of authors, the frequencies of punctuations were used as in [32, 33].The proportion of plural and singular nouns, pronouns and proper nouns, the ratio of past and future verb tenses, ratios of comparative and superlative adjectives and adverbs were used by many researchers.

E. **Structural Features**

Structural features appear in managing the organization of text and its outline called structure. These features have demonstrated especially imperative in analysing the web messages as in [26]. The inscription style of an author is identified by both the features of style mentioned earlier and the structural information related to the paragraph. Researchers usually concentrates on structure of the words such as good wishes signatures and on the total count of paragraphs and average length of the paragraph, number of special characters, sentence length, words per sentence and the style of writing lengthy complex sentences are the features which contribute for identification of style. Though these features are significant discriminators, they don't hold the extra information enclosed in web messages. Another new category of features associated with structural information named technical structure to envelop textual style, hyperlink and entrenched image characteristics are addressed in the work of [25]. The researchers has to have knowledge about the length of the conversation, the presence of hyperlinks, images and the style used either at the beginning or at the end of the conversation. It is addressed that Conversation length as feature is useful in spam detection by Michał Meina. It is observed from the literature that generally the higher age people use longer words with greater frequency and females wrote longer sentences than males. Average sentence length used by [34], [35], the number of HTML tags were used by [36], the number of URL's were used by [33], [34], [36], the set of common slang vocabulary were used by [37], [38] and the number of emoticons were used by [36], [39], [33], [40], [41], [42], [38], [43].

F. **Content-Specific Features**

In a Particular domain topic, a specific set of words will come on a regular basis those words are called Content-specific features .While discussing about computers some words like RAM ROM, LAPTOP, DESKTOP will appear these words are treated as content specific features. Feature set of English language incorporated 301 features altogether, with syntactic (158), content-specific features (11), lexical (87), structural (45) in [26].

IV. APPROACHES OF AUTHORSHIP ATTRIBUTION

The techniques proposed on attribution in the beginning period were PC-helped rather than PC-based. Most successful and popular at that time was CUSUM or QSUM .Majority of the researchers not agree with CUSUM . Stamatatos [15] utilized 1000 most regular words as features. Word N-gram are the lexical features used . These can be extracted by using character n-grams by putting n=1, n=2 and n=3. Baayen et.al [17] used other approach which is based on semantic dependency graphs which explains semantic dependency between words, and McCarthy et.al. used Coh-Metrix in [44]. Any authorship-attribution process consisting of a list of candidate authors , a collection of documents of all candidate authors whose authorship is known and a group of documents of unknown authorship, everyone them must be attributed to a candidate author. The study of stamatatos et.al plainly distinguished the approaches of authorship attribution according to whether they consider every training document independently or collectively. The investigation specified a way to deal with link all the accessible training texts per author in one large file and collective depiction of author's style know as author's profile from the combined text. This is called profile based approach [11]. The alternative approaches need several training samples of each author to build up an accurate attribution model. That means, every training text is independently represent a distinct case of author style. It is called as Instance-Based approaches. The modern authorship-attribution approaches believe each training text sample as a single separate piece which contributes independently to attribution model. That is every sample of known authorship is an instance of a problem in query.

Every training sample of a text in the corpus is represented by a attribute vector, a trained classification algorithm with the set of instances of known authorship is used to construct an attribution model. Then this model finds the correct author of the unidentified text. For this type of approaches classification algorithms involve many training instances per class to obtain a reliable model. So, in instance-based approaches, in the event that we have just a single, however a very long, training text for a specific candidate author (e.g., an entire book), this ought to be splitted into equal length numerous parts.

Size wise normalized text for training is used whenever there are no. of text sample of varying length, for a single author. Samples of Equally portioned training texts of each author are used (Sanderson & Guenter, 2006). In each of these cases, the samples has to be sufficiently long with the goal that the features can present to their style. Different lengths of text samples have been accounted here in the survey. Sanderson et.al and Koppel et al. in [45] has grouped the text of size consisting of 500 words. Feiguina et.al carried investigations with text chunks of different length (i.e., 1000, 500,200, words) and concluded considerably diminished accuracy as the text-chunks length diminished. so, the selection of the training text sample is not a significant process and straightforwardly influence the performance of the attribution model.

In profile based approaches the text is represented as single file which includes the training text of all authors. In Instance-based approaches a separate file s required for representing a single author. It is very difficult to

combine different features in Profile-based approaches where as in Instance-based approaches. features can be combined easily. In Profile-based approaches classification models like Bayesian and similarity based methods used where as in Instance-based approaches Discriminative models, Powerful machine learning algorithms (SVM) are used. Profile-based approaches require very less time to train where as Instance-based an approach requires relatively high time.

V. METHODS OF AUTHORSHIP ATTRIBUTION

AA Strategies are partitioned to three fundamental classes. Machine learning techniques that are regularly used text categorization. In this technique the known writing of every author is used to build a classifier that can be utilized to classify unique texts.SVM and NN are the good examples of this techniques.

A. **Machine Learning Methods**

The study in machine-learning technique is always concentrates on the selection of features in representation of document and on the selection of learning algorithms. Methods of selection are relay on whether there are two or more candidate authors. If only two candidate authors exists, then use Support Vector Machines SVMs are better descriptions of instance-based approaches are done by the vector space models. These algorithms were studied thoroughly in the area of topic-based text-categorization investigations (Sebastiani, 2002) in [8]. A few of these algorithms can efficiently manage multi-dimensional, noisy, and sparse data, permitting significantly numerous ways of presenting the texts. For an instance, whenever several features are used, an Support Vector Machine model is capable to avoid over fitting problems and is viewed as one of the finest solution of present technology as in [4], [11].Class-imbalance is problem which is effecting the vector space model. A new technique was proposed by Stamatatos et.al. to handle this type of problems with the use of instance based approaches. Training set text samples can be segmented according to the size of their class. In this way several small text samples are prepared for minority authors (authors with less no. of for training sample) while few, but lengthy, texts can be prepared for majority authors (the authors with multiple training texts).

B. Similarity Based Methods

Distance methods also called similarity based methods that determine the likeness among feature vectors depending upon a distance formula. The major thought of distance-based methods is the computation of pair wise similarity measures among unseen text and all the training texts, and then based on a nearest-neighbour algorithm the estimation of the most likely author is found. In this type a suitable measure is applied to measure the distance between two documents, and an unidentified document is assigned to that author to whom the document is much related. The Study of distance-based technique's concentrates on the selection of features for text representation, process for dimensionality reduction like Principal Components Analysis of features, and on the selection of similarity measurement. In the study of Fakotakis [46] the usage of mahalonobi distance and J. Savoy et.al used mean difference of z-score weight's among unattributed text and training text. If there are many, follow Koppel approach, which is denoted as KOP. The principal aim of KOP is to group the authors into pairs and they are discriminated by dissimilar subsets of the feature space. So, KOP arbitrarily selects k1 subsets of length k2 from a list of F features; then for all of these k1 subsets, KOP calculates the cosine similarity among a test sample text and all the text documents by single author. Kop then returns the author who had the majority of the best matches. Higher precision cases can be managed by putting a threshold value in Kop w.r.t. minimal recall. When top match is less than threshold value then KOP returns "unknown author".

The vital approach of this classification was proposed by Burrows et al., [10] and named it as "Delta". This strategy measures the z- distribution of a group of function words. Then, for every document, the deviation of each word frequency is evaluated in terms of z- score. Then KOP check's whether word is used more or less times than the average no. of times .If it is more the z-score is positive otherwise z-core is negative. Ultimately, the Delta measure showing the contrast between an arrangement of (training)text composed by a similar author and an unknown text is the mean of the absolute differences between the z scores for the entire function word set in the training texts and the corresponding z scores of the unknown text. If the Delta measure is very less then similarity between the attributing text and the candidate author is very high. Hoover, 2004a et.al. in [11] proved that literary texts like novels, poems are most suitable for this method and produced significant results. Argamon (2008) in [47] explained the working principle of Delta theoretically and demonstrated that Delta can be seen as a axis-weighted form of nearest neighbour classification where the unidentified text is allocate to the nearest class instead of the nearest training text. Hoover et.al. in [11] worked on thorough research of disparity of Delta. Also concluded that increased accuracy of delta is achieved when bigger sets of frequent words of size 500 an above. Benedetto et al. (2002) depicted a new distance-based approach which uses text compression models to approximate the dissimilarity between texts. The learning part of this method purely includes the compression of every training text in different files using GZIP algorithm. In identifying the authors of an

unidentified text, this text is appended to every training text file, and then each resultant file is compressed by the same algorithm.

C. Meta-Learning Methods

Alternate to normal classification algorithms reported earlier numerous much complex algorithms were exclusively developed for authorship attribution. In this current category existing classification algorithm could serve as a tool in a meta-learning technique. Koppel et al. 2007 in [45] proposed a much attractive approach called unmasking method. It is generally used for authorship verification. The major disparity with the classical instance-based approach and unmasking method is that the training phase is not required. For every unattributed text, an SVM classifier is constructed to distinguish it from the training texts of each author. Therefore, Koppel et al. constructed N-classifiers for N- authors for every unattributed text.

D. Hybrid Approaches

Van Halteren et.al. (2007) in [48] Explained an approach that utilizes a few basics from both instance-based and profile based approaches called hybrid approach. All the training text samples were represented independently similar to the instance based approaches. Every text of each author is represented by a vector which is feature-wise averaged and formed as one unique profile vector. This is the similar case with profile-based approach. Similarity of the profile of an unattributed text from the profile of every author was measured as a weighted feature-wise function. Grieve (2007) et.al. in [20] build a equivalent hybrid approach.

E. Probabilistic Methods

Probabilistic Methods finds candidate author A, given that maximize the probability of P (A | U) for unattributed text U. naive base is the widely used probability based classifiers [49].Topic models in [50] and language modeling are the other models. Probabilistic language model is the language model basic model in Information retrieval which constructs a Probabilistic model is an information retrieval model which constructs the language model for every document and arranges the document depending on the structure of a query. Initially the probability model was build for the training documents of all authors and then the actual author was attributed based on the highest probability of occurrence from the unknown text.

VI. MODELS OF AUTHORSHIP ATTRIBUTION

Various models were examined depending up on the correctness of various features like sample size, count of authors and the diffusion of training texts between the authors as in [7].

A. **LDA Model**

LDA model is a topic identification model wherein it contains a symmetrical Dirichilet Distribution and there are 3 parameters which are generated by choosing topic from a document, choosing a token from the token topic distribution. This model was inferred from the data using Gibb's sampling approach. It is observed that the topics obtained by LDA may not comparable to that of human interpretable topics. In the literature latent factors was introduced using two way utilization of LDA in author attribution. These two models are LDA with topic SVM and LDA with Hellinger. In both methods the role of LDA is to applying a frequency filter on to the exiting for dimensionality reduction.

B. **Topic SVM**

The binary SVM classifier is used to discriminate the authors who are identified based on the topic distributions used as features. Similar approach was used for document classification but it was adopted for Author Attribution by considering stop words only.

C. **LDA with Hellinger**

In this approach the most likely author of a document is identified by finding the distance between the document topic distributions which consists of many candidate authors.

D. **Multi Document LDAH-M**

In this model possible author is identified by lowest mean distance for all his documents wherein the distance is measured from Hellinger distance model wherein the distribution is among the training documents.

VII. CONCLUSION

Various methods, classifiers and approaches were use and presented in this paper. This paper aims at identifying the suitable model and there by analyzing the disadvantages of models in order to propose a new model. Various machine learning algorithms were studied and witnessed the applicability of these machine leaning techniques

onto Authorship attribution was presented. A proposal is made to identify the least important features from the repository and applicability of usable features is planned as a new proposal for Authorship attribution.

REFERENCES

[1]. Raghunadha Reddy T, Vishnu Vardhan B, Vijayapal Reddy P, "A Survey on Author Profiling Techniques", International Journal of Applied Engineering Research, March 2016, Volume-11, Issue-5, pp. 3092-3102.

[2]. M. Sreenivas, Raghunadha Reddy T, Vishnu Vardhan B, "A Novel Document Representation Approach for Authorship Attribution", International Journal of Intelligent Engineering and Systems, 11 (3), pp. 261-270, MAY 2018.

[3]. P. Jeevan Kumar, G. Srikanth Reddy, T. Raghunadha Reddy, " Document Weighted Approach for Authorship Attribution" in International Journal of Computational Intelligence Research, Volume 13, Number 7, pp. 1653-1661, 2017.

[4]. Zheng, R., Li, J., Chen, H., & Huang, Z. : "A framework for authorship identification of online messages: Writing style features and classification techniques". Journal of the American Society of Information Science and Technology, 57(3), 378-393, (2006).

[5]. Juola, P. ,"Authoship Attribution. Hanover, MA: Now Publishers" (2008).

[6]. Koppel , M., Schler, J., & Argamon, S., "Computational Methods in Authorship Attribution. Journal of the American Society for Information Science and Technology", 60(1), 9-26. (2009).

[7]. Koppel, M., Schler, J., & Messeri, E. , " Authorship Attribution in Law Enforcement Scenarios. In C.S. Gal, P. Kantor, & B. Saphira (Eds.), Security Informatics and Terrorism: Patrolling the Web" (pp.111-119). Amsterdam: IOS, (2008).

[8]. Abbasi, A., & Chen, H., "Writeprints: A Stylometric Approach to Identity-Level Identification and similarity Detection". ACM Transactions on Information Systems, 26(2), 1-29, (2008).

[9]. Stamatatos, E., Fakotakis, N., & Kokkinakis : "Computer-based authorship attribution without lexical measures" . Computers and the Humanities, 35(2), 193-214, , G. (2001).

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

[11]. Hoover, D.: "Testing Burrows Delta". Literary and Linguistic Computing, 19(4), 453-475. Hoover, D. 2004a

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

[13]. Chaski, C.E.: "Who's at the keyboard Authorship attribution in digital evidence investigations". International Journal of Digital Evidence, 4(1), (2005).

[14]. Grant, T. D.: "Quantifying evidence for forensic authorship analysis". International Journal of Speech Language and the Law, 14(1), 1 -25, (2007).

[15]. Frantzeskou, G., Stamatatos, E., Gritzalis, S., & Katsikas: "Effective identification of source code authorusing byte-level information". In Proceedings of the 28th International Conference on Software Engineering (pp. 893-896), (2006).

[16]. Argamon, S., Šarić, M., & Stein, "Style Mining of Electronic Messages for Multiple Authorship Discrimination:" First Results. Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (2003).

[17]. Gamon, M "Linguistic Correlates of Style: Authorship Classification with Deep Linguistic Analysis Features", Proceedings of the 20th International Conference on Computational Linguistics: Vol.4 (pp. 611-617). (2004). Stroudsburg, PA: Association for Computational Linguistics.

[18]. Juola, P., & Baayen, H "A Controlled –Corpus Experiment in Authorship Attribution by Cross-Entropy". Literary and Linguistic Computing, 20(1), 59-67,. (2005).

[19]. Peng, F., Schuurmans, D., Keselj, V., & Wang, S.,"Language Independent Authorship Attribution Using Character Level Language Models". Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics: Vol. 1 (pp. 267-274). Stroudsburg, PA: Association for Computational Linguistics, (2003).

[20]. Grieve. J., "Quantitative Authorship Attribution: An Evaluation of Techniques". Literary and Linguistic Computing, 22(3), 425-442, (2007).

[21]. Hirst, G., & Feiguina, O, "Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts". Literary and Linguistic Computing, 22(4), 405-417, .. (2007).

[22]. Raghavan, S., Kovashka, A., & Mooney, "Authorship Attribution Using Probabilistic Context-Free Grammars". Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (pp. 38-42), (2010).

[23]. Rico-Sulayes, A., "Statistical Authorship Attribution of Mexican Drug Trafficking Online Forum Posts." International Journal of Speech, Language and the Law, 18(1), 53-74, (2011).

[24]. Raghunadha Reddy T, Vishnu Vardhan B, GopiChand M, Karunakar K, "Gender prediction in Author Profiling using ReliefF Feature Se-lection Algorithm", Proceedings in Advances in Intelligent Systems and Computing, Volume 695, PP. 169-176, 2018.

[25]. M. Sreenivas, Raghunadha Reddy T, Vishnu Vardhan B, "Author Identification Using Information Retrieval Features", International Journal of Computer Engineering and Applications, Volume XII, Issue IV, pp. 22 – 27, APR 2018.

[26]. O. De Vel et al.: "Mining E-mail Content for Author Identification Forensics," SIGMOD Record, vol. 30, no. 4, 2001, pp. 55–64.

[27]. Christopher Ian Baker, "Proof of Concept Framework for Prediction", Proceedings of CLEF 2014 Evaluation Labs, 2014.

[28]. Magdalena Jankowska, Vlado Kešelj, and Evangelos Milios, "CNG text classification for Authorship Profiling task", Proceedings of CLEF 2013 Evaluation Labs, 2013.

[29]. Magdalena Jankowska, Vlado Kešelj, and Evangelos Milios, "CNG text classification for Authorship Profiling task", Proceedings of CLEF 2013 Evaluation Labs, 2013.

[30]. Khmelev, D.V., & Teahan, W.J : "A repetition based measure for verification of text collections and for text categorization". In Proceedings of the 26th ACM SIGIR, (pp. 104–110),. (2003a).

[31]. P. Buddha Reddy, T. Raghunadha Reddy, M. Gopi Chand and A. Venkannababu, "A New Approach for Authorship Attribution", Advances in Intelligent Systems and Computing, vol. 701, pp. 1-9, 2018.

[32]. K Santosh, Romil Bansal, Mihir Shekhar, and Vasudeva Varma, "Author Profiling: Predicting Age and Gender from Blogs", Proceedings of CLEF 2013 Evaluation Labs, 2013.

[33]. Yuridiana Aleman, Nahun Loya, Darnes Vilarino, David Pinto, "Two methodologies applied to the Author Profiling task", Proceedings of CLEF 2013 Evaluation Labs, 2013.

[34]. Wee-Yong Lim, Jonathan Goh and Vrizlynn L. L. Thing. "Content-centric age and gender profiling", Proceedings of CLEF 2013 Evaluation Labs, 2013.

[35]. Braja GopalPatra, Somnath Banerjee, Dipankar Das, Tanik Saikh, Sivaji andyopadhya, "Automatic.Author Profiling Based on Linguistic and Stylistic Features", Proceedings of CLEF 2013 Evaluation Labs, 2013.

[36]. Raghunadha Reddy T, Vishnu Vardhan B, Vijayapal Reddy P, "Author profile prediction using pivoted unique term normalization", Indian Journal of Science and Technology, Vol 9, Issue 46, Dec 2016.
[37].
[38]. A. Pastor López-Monroy, Manuel Montes-y-Gómez, Hugo Jair Escalante, Luis Villaseñor-Pineda, and Esaú Villatoro-Tello, "INAOE's participation at PAN'13: Author Profiling task", Proceedings of CLEF 2013 Evaluation Labs, 2013.
[39]. 9Raghunadha Reddy T, Vishnu Vardhan B, Vijayapal Reddy P, "Profile specific Document Weighted approach using a New Term Weighting Measure for Author Profiling ", International Journal of Intelligent Engineering and Systems, 9 (4), pp. 136-146, Nov 2016.
[40]. Suraj Maharjan and Thamar Solorio, "UsingWide Range of Features for Author Profiling", Proceedings of CLEF 2015 Evaluation Labs, 2015.
[41]. Raghunadha Reddy T, Vishnu Vardhan B, Vijayapal Reddy P, "N-gram approach for Gender Prediction",Â 7 th IEEE International Advanced Computing Conference, Hyderabad, Telangana, PP. 860-865, Jan 5-7, 2017.
[42]. Lucie Flekova and Iryna Gurevych, "CanWe Hide in theWeb? Large Scale Simultaneous Age and GenderAuthor Profiling in Social Media", Proceedings of CLEF 2013 Evaluation Labs, 2013.
[43]. Miguel A. Álvarez-Carmona, A. Pastor López- Monroy,Manuel Montes-y-Gómez, Luis Villaseñor- Pineda, and Hugo Jair Escalante, "INAOE's participation at PAN'15: Author Profiling task", Proceedings of CLEF 2014 Evaluation Labs, 2015.
[44]. Lesly Miculicich Werlen., "Statistical Learning Methods for Profiling Analysis", Proceedings of CLEF 2015 Evaluation Labs, 2015.
[45]. Raghunadha Reddy T, Vishnu Vardhan B, Vijayapal Reddy P, "A Document Weighted Approach for Gender and Age Prediction", International Journal of Engineering -Transactions B: Applications, Volume 30, Number 5, pp. 647-653, May 2017. Koppel, M., Schler, J.,&Bonchek-Dokow, E., "Measuring differentiability: Unmasking pseudonymousauthors". Journal of Machine Learning Research, 8, 1261–1276, (2007).
[46]. Stamatatos, N. Fakotakis and G. Kokkinakis , "Automatic Text Categorization in Terms of Genre and Author," Computational linguistics , vol. 26, no. 4, pp. 471-495, 2000.
[47]. Argamon, S. "Interpreting Burrows' Delta: Geometric and probabilistic foundations". Literary and Linguistic Computing, 23(2), 131–147, (2008).
[48]. Van Halteren, H., "Author verification by linguistic profiling:An exploration of the parameter space". ACMTransactions on Speech and Language Processing, 4(1), 1–17, (2007).
[49]. F.Peng and D. Schuurmans., "Combining naive Bayes and n-gram language models for text classification" 2003.
[50]. Y. Seroussi, I. Zukerman and F. Bohnert."Authorship attribution with topic models," Computational Linguistics, vol. 40, no. 2, pp. 269--310, 2014.

Sreenivas Mekala "A Survey On Authorship Attribution Approaches "International Journal of Computational Engineering Research (IJCER), vol. 08, no. 09, 2018, pp 48-55