International Journal of Electrical and Computer Engineering (IJECE)
Vol. 9, No. 4, August 2019, pp. 3167~3174
ISSN: 2088-8708, DOI: 10.11591/ijece.v9i4.pp3167-3174

A robust authorship attribution on big period

Mubin Shoukat Tamboli[1], Rajesh Prasad[2]
[1]Matoshri College of Engineering and Research Centre, India
[2]Sinhgad Institute of Technology and Science Narhe, India

Article Info

Article history:

Received Aug 26, 2018
Revised Mar 13, 2019
Accepted Apr 4, 2019

Keywords:

Author identification
Attribution
Feature extraction
Stylometry
SVM

ABSTRACT

Authorship attribution is a task to identify the writer of unknown text and categorize it to known writer. Writing style of each author is distinct and can be used for the discrimination. There are different parameters responsible for rectifying such changes. When the writing samples collected for an author when it belongs to small period, it can participate efficiently for identification of unknown sample. In this paper author identification problem considered where writing sample is not available on the same time period. Such evidences collected over long period of time. And character n-gram, word n-gram and pos n-gram features used to build the model. As they are contributing towards style of writer in terms of content as well as statistic characteristic of writing style. We applied support vector machine algorithm for classification. Effective results and outcome came out from the experiments. While discriminating among multiple authors, corpus selection and construction were the most tedious task which was implemented effectively. It is observed that accuracy varied on feature type. Word and character n-gram have shown good accuracy than PoS n-gram.

Corresponding Author:

Mubin Shoukat Tamboli,
Department of Computer Engineering,
Matoshri College of Engineering and Research Centre,
Nashik, Maharashtra, India.
Email: mubin.tamboli@gmail.com

1. INTRODUCTION

Authorship identification is distinguishing task of recognizing writer of a given content from his writing style. Research of author identification has got exponential growth in recent years due to its valuable contribution in forensic, linguistic research, social psychology, literary science, social media analysis and e-commerce activities. Due to the bloom of Internet communication has become easier and common turn was diverted into venomous movements. We identify such suspicious entities over the network. Actually, this is a tedious task but it can be simplified using authorship attribution. Generally, messages on web are nameless. Many authors in their writing, they don't give their genuine character data. For example, name, age, sex and address. In numerous abuses or wrongdoing instances of online messages, it is required to find the real identity of authors. Along these lines, the obscurity of online messages forces some kind of difficulties to identify author of contents available on Internet. In the nature of attribution, categorization is made for unknown text document to one, from a limited set of candidate author whose data-set is in terms of writing sample [1].

Writing of each author is uniquely identified with writing style. So, writing style would become one metric of discriminating authors. There are factors affect the performance of attribution system: number of authors, writing sample, size of writing sample, period of known and unknown writing samples available. Writing style of author changes over time. To observe such change and its effect on author identification, dataset should contain sample over a big time period, so that performance of system can be evaluated correctly. There are several parameters which affects writing style of individual when big period considered.

These factors effects on style of writing includes education, nationality, genre, age, topic, formal content, written in the same period. These samples are written over different time period, cognitive distortion, thinking, emotions, psychology etc. These challenges have once in a while been tended to by the research group. Many techniques have shown remarkable adequacy in distinguishing the genuine writers, and in review several researches made by considering sample in same time period hence distortion in writing style considered negligible [2]. Artificial neural network can be used by text processing [3]. There is remarkable change in writing style at different age at composition novel, conclude that vocabulary size limited over time [4]. Temporal changes occurred in writing style and vocabulary usage on short text over time. Author do have change in his writing style but different authors have different style [5]. Time and topics are responsible for change in writing styles and also affect the accuracy of attribution task [6-7].

In traditional approaches, there are some disadvantages. First is that writing sample used in the attribution task are irrespective of time period. Probably all samples belong to near periods. So, the time effect on every sample is nullified. Second is that research limited to fixed number of features worked on limited to similarity-based approach for content types of features and document statistic features uses machine learning approaches. Azarbonyad H. [8] focused on author identification with the consideration of time period but approach time period limited to four years and used feature type is character 4 gram.

In this paper, we found accuracy affected on identification, for document sample collected over big time. We accumulate dataset from news articles and collected letters from famous personalities. Formulate new approach for author identification from the inspiration of time-based language models [5, 9, 10]. Dataset consist of writing sample written by author over and average thirty years of time period. Features, a selective set of attributes on which machine learning approach is followed. Our methodology worked on features as character sequencing, word sequencing, part of speech sequencing and combination of them. This is followed by machine learning approach for classification. SVM supervised machine learning approach used for distinguishing the unknown sample to make it known. In described approach, a Bag of Word (BoW) representation strategy applied. Simplified corpus structure is used, a raw text samples distinguished over time space with the identification of month and year in which document written. Remaining paper is organized as below. Section 2 describes the related work done in the area. Methodology described in section 3. Section 4 elaborates experiment results and discussion and conclusion be the last section 5 of this paper.

2. RELATED STUDY

Authorship attribution works in three main domains- author identification, similarity detection and characterization. In identification task, the history of authorship style is known in advance and likelihood of writing sample with available information. Author characterization outlines the attributes of a writer and produces the writer profile in light of his or her work. Some of these qualities include sex, instructive and social foundation and dialect commonality. In similarity detection, work of different authors is compared with single author to find it its closeness [11]. Detailed review [7] describes the methodologies for authorship attribution starting from stylometry features, which consist of lexical features, character features, syntactic feature, semantic features, application specific features. Lexical highlights are words or character-based factual measures of a lexical variety [12]. Character based features consist of sequence of characters, measured at character level. Syntactic feature involves syntax of language used for writing and semantic corresponds to meaning of sentence for which NLP tools can be used, application type involves specifically type of document, domain of discussion in corpus [11]. Features are nothing but style markers. Paper [13] uses a similarity-based approach with random features, which captures information about topic and writing style and applied novel algorithm to produce result. Result by them was observed on about 1000 author with 93.2% precision. Focus was given only for character 4-gram, a single feature with about 20000 attributes. Then this result was compared with distributing dataset into training and testing, repeatedly.

Research in visualized event driven approach [14] can be visualized and interpreted. Approach was based on visualization of fingerprint comparison. Feature set is based on two types, set one is unified and another is class specific. A group of writing style feature gathered as evidence unit along with its scoring vector. Analysis was made on such event score to find targeted creator. Suitability of method is limited to 20 authors. Word n-gram, character n-gram, PoS n-gram features used to construct event. The approach explained in [15], based on theoretical study of function words in authorship attribution. Function words are nothing but extraneous words used without affecting meaning of sentence. Stamatos [16] elaborated fast test categorization methods. In the work, used NLP tool to eradicate the stylistic facts. Multiple regression and discriminant analysis classification models were used to categorize creator's facts. One of the research paper [17] deals with source code written by different programmer which was identified on the basis on n-gram author profile. This method is based on byte level n-gram features on different source code written by

different author in java or C++ language. In this approach, fixed number of most frequent n-gram from source code file was considered.

A similarity measure defined by Keseji [18] was used to find relative distance among two styles. On the available dataset, the approach shows 88% accuracy. Spatium-L1, an effective author verification model based a structure of unsupervised learning algorithm [19]. It makes use of 200 mostly occurred terms of unknown text. The method described in the paper was distance based. Author has represented his work in PAN CLEF 2014 and gives good comparative result against existing algorithms. Used features in the methods were word type and punctuation symbols. A case study on vocabulary changes [4] on lexical, syntactic and discourse was observed. Corpus from novel was used for experiments and only first 50000 words used. Observed facts were count of vocabulary size, richness and unique words as a one measure and word n-gram, word length was other and third type as occurrences of vague and indefinite words were assumed as facts. At different age of composition found variations on considered facts. Another challenging task in authorship attribution is to handle messages with varying length. Digital communication over internet is always in form of short messages in terms of email, chat messages, tweets etc.

Zheng [11] comes with one of the idea to identify author for online messages. English and Chinese language chosen for experimenting. Features were considered in groups as lexical features, word-based features, syntactic features, content-specific features, structural features, C4.5, NN, SVM classification models used for identification. Results based on feature types and techniques like SVM and NN produced challenging results. Cognitive error is another aspect for attribution. A writer makes basic mistakes with regard to few issues specifically: Causal Premise, Probability Judgment and Conditional Reliance. The researcher's blunders are imperative since they have a large effect on his conclusions and since comparative mistakes frequently happen when individuals, both specialists and amateurs are confronted with the challenges of Bayesian inference [20]. Cheng [21] recognizes that the issue of sexual orientation recognizable proof from text is an interaction between psycho-linguistics, nonspecific writing styles of men and ladies. In their studies they have used three algorithms viz SVM, Bayesian logistic regression and Adaboost decision tree. Accuracy captured is around 85%. Contributed features in discrimination were function words, word-based features, structural features. The model applied on Corpus from reuters and enron email dataset. Text is represented in the form of vectors. Many types of window algorithms are applied to discriminate among several authors [22] to produce compromising accuracy. Writeprint, a new technique introduced [12] in which sliding window features were considered for the application of language model. Many types of features were accumulated and applied on this new model which produces accuracy around greater than 90%. All these reviewed researches focused on the methods for author attribution and identification, time when the document generated was not considered.

Azarbonyad [8] studied attribution where writing nature of author changes. These temporal changes are observed with respect to word distribution in writing samples. In the experiment done, tweet and enron email data set was considered over the period of 5 years. Character 4-gram was the observed features. Temporal changes were captured with algorithm defined in paper [9, 10], time-based language model and calculated from linear regression techniques. Research work [4] elaborates change in vocabulary usage by a writer and proved that size of vocabulary goes on decreasing over time. Time frame over the work was big about up to 35 years. Author [6] also investigated and concluded writing style of author changes over time and authorship verification accuracy increases when record is composed in a brief period of time. Review of different methods of classification and their results is elaborated in paper [2]. And [23] uses machine learning approach for content types of feature where writing samples are shorts and irrespective of time. In [24] show the variation of features over time. Variation were not stationary in the work. But indicates there is change in writing style of author. In the survey [25], describes different types of features and their combination and applied on arabic text for to the authorship attribution task.

2.1. Feature selection

Authorship attribution is an information retrieval task, where feature used in the work can affect the outcomes produced. From above literature, features are broadly categorized in types such as writing style and content specific. In writing style, writing style of author is captured in different means as style of author. In content specific, importance is given to writing content and its meaning. In content specific features, we are mostly concerned about word n-gram, character n-gram, frequency of word usage, cognitive errors made, content specific features etc. These features are considered as group because each type of the feature is consisting of many attributes. It is always evaluated in groups. Sometime error occurred in document can also be treated as one of the effective features. There are features which captures both aspect content and style, character n-gram is one of them.

2.2. Character n-gram

In linguistic arithmetic, character n-gram is consisting of a continual sequence of n terms in a given sample. It can be phoneme which is a gesture of sound which differentiates one from another, syllable which distinguishes sequence of speech sound in words e.g. mathematics composed of two syllables viz mathe and matics, letters, words etc. In continual sequence, it is needed to define whether it is in word or complete document. It means whether it considers the space, operator, and punctuation in words or not.

2.3. Word n-gram

In this context, text is viewed as sequence of words. In word n-gram words are collected, which has n contiguous words. It captures content specific information rather than stylistic information from the corpus [7]. According to survey made by Stamatatos, word n-gram is used for author identification. It doesn't always give promising results rather than others. Many times, such contiguous words are not always occurring in system, specifically short text. It may not give correct information all the time because it is incapable when writing error is introduced in sample. It captures human behavior but it is possible that the behavior may change over time. For short text, there is less possibility of capturing such repeating behavior. In paper [23] word n-gram feature is used to gather semantically meaningful information from sample of short text.

2.4. Part of speech n-gram

This feature captures stylistic information from given sample of text. To generate this feature, first it is required to tag the text sample. It is the way of increasing a word in content (corpus) as comparing to a specific grammatical form, in light of the two. Its definition and relationship with nearby word, sentence and paragraph. It is firmly fixing to corpus etymological. Universal tag set consist of following tags.

2.5. Function words

Function words are wording whose design is to add the linguistic structure as opposed to the significance of a sentence. It is open class word, which includes adjective, noun, verb etc. Example of function words are 'of', 'at', 'in', 'that', 'do' etc. Function words can be considered as a base for textual comparison. In general count of function words in any sample is large which can act as distinguished feature for identifying author. It is not focusing towards the content of text; it focuses on stylistic feature of an author. It cannot be directly applied to discriminate between authors. In [26] methodology have more than 175 function words.

2.6. Classification

Authorship verification problem can be solved by similarity-based approach and machine learning approach. Machine learning classification model follows instance-based approach. In instance-based approach each training sample is identified uniquely in attribution model. Classification model uses various supervised and unsupervised method for the attribution. In the description we obey SVM machine learning method as instance-based approach for classification [7], [23], [19]. In machine learning based approach, writing style of each known author is identified as training sample, which is used to build classifier, which on next used to classify unknown sample. Here is need of enough sample to train classifier so it can be used on further [13]. Decision tree algorithm [27] used in author identification for Marathi language. Similarity based approach is another one, here distance metric is used to discriminate between two sample, if one sample is most similar with another then conclude that is written with same author. This is direct approach hence not considered in an instance-based approach.

Support Vector Machine act as discriminating classifier which separates data sample through a hyperplane. A number of hyperplanes separate out data sample into number of classes. It is suitable to work with high dimensional data, hence suitable for our approach. It consists of set of training point act as vector which is represented in bag of words [28] form. Various kernel functions can be used, it is a kind of algorithm used for pattern analysis. A kernel is used to make linear model to nonlinear model. SVM supports for different type of kernel [29].

3. METHODOLOGY

Our approach is based on representation of sampled text token and their groupings. Most similar author for the identification based the classification made by our system. We followed machine learning approach to build model. Features representation in terms of vector and used instance base approach and followed by SVM algorithm for constructing model. Different feature sets used to setup model, for verification of the model we define one set of samples as training and one set as testing. And from this we

decide the effectiveness of our algorithm. Proposed methodology builds with feature representation, feature selection, application of classification algorithm to build model and then verifying results. In the feature selection, initially we remove out all the special symbols, URLs, function words from the text. Function words removed as it is commonly used repeating terms, and in the procedure extraction of style is the primary objective. While extracting character, word and pos sequence are respective to each sentence. All the extracted features are represented in the bag of words. Rest of the procedure is listed in the algorithm given below.

Given:
n number of sample from which *k* sample act as training sample and *n-k* act as testing samples.
Algorithm:
V = feature Vector in terms of Bag of Words representation (v_k *U* v_{n-k})
t = { *char-n-gram, word-n-gram, pos-n-gram* }
Repeat for each sample S_i
Text = Preprocessing*(sample)*
V = *BuildFeature*(feature type t_i, *Text*)
For each sample feature from V
$V_{new,k}$, $V_{new,\ n-k}$ = ReduceFeatureDimension(V)
R = BuildClassifierSVM(PolynomialKernel(*pow 2*), V_k,V_{n-k})
For each sample *s* from set of testing sample
best match for *s* is derived from R
Output:
Each sample from testing set is assigned to best match class, number of classes are same as author set.

The main idea with above methodology is that it works in good way for text sample which are collected from user over long period, it proves the algorithm produces effective results. In our Bag of Word sample is a dictionary-based representation of data. In Feature extraction methodology, we used NLTK tools to extract character word, PoS n-grams. We used sklearn library to represent our data in bag of words formatted vector. All these features combined together with concatenation operation to from alone feature vector. In the section of normalization, we normalized the sample among the whole feature vector.

4. EXPERIMENT RESULTS AND DISCUSSION

We have collected numerous English language data from real world entity. We have collected digital as well as handwritten docs whose digital copy is available online. We have collected data of 11 authors of various time span. The period in which data is available from 4 years to more than 30 years. Mix datasets were considered. In our dataset, mixed time period corpus is used. From available resources, we accumulate handwritten letters of few authors which converted to digitized form. And on further used in identification process. Out of eleven authors, six authors corpus was from time period 5 to 8 and five author having corpus from time period 25 to 38. An accumulated corpus size about more than 300 words in a document.

From text sample we first removed special symbols, and repeated words as in format of letter were removed. ASCII characters kept as it is in samples. All text sample not from same time period so our system becomes robust. We use bag of words format for representation of our feature vector, as used features were character n gram, word n gram and pos n gram. We utilized all such feature uniquely.We checked comparative result for the same. When we constructed feature vector, we select k best feature from them and applied classification model to generate result. Selected feature evaluated on the basis of how correctly the identification of author made. Following Table 1 illustrate how correctly the author identified.

Table 1. The performance of authorship attribution

Feature	Time span	Accuracy	Kappa Statistics
POS2gram		78.65	0.76
POS4gram		70.17	0.67
Word2gram		84.21	0.82
Word3gram	Average 30	70.46	0.67
Word4gram	years	50.87	0.46
Char4gram		86.98	0.86
Char5gram		88.59	0.87
POS2Wordd2		83.33	0.81

On use of selected feature type indicated in Table 1, the result varies from 50% to 88%. When talk about POS as features, result is good but not enough, it shows for pos2gram 78% and pos4gram 70%, means as n increases for POS-n-gram performance of classifier degraded. In word-n-gram word2gram gives good accuracy as two continuous word occurrence is good characteristic to identify author habit than any other feature. It makes sense for selecting feature as compared to others. For character 4 and 5 gram indicates close result, most of the research work suggest the character 4 gram are good feature to capture users writing style as well content used in corpus. But it makes no sense, how it gets capture both, if used, we identified habitual mistake, repetition of writing, as words broken into n character, it doesn't have any mean. Though such problems arise in character n gram, it gives promising result than any other features.

We also tried for combined feature type as pos-2-gram and word- 2-gram combinedly, but with this, feature size increases and hence execution time for the system increases. But results are effective but as compared to word-2-gram it is less. Table 1 shows the accuracy and kappa statistic for various feature type. Kappa statistic used to measure interrater reliability. When statistic is greater than 0.8 then agreement level strong and if less than 0.6 then it is week. In Table 1, word-2-gram, char-4-gram, char-5-gram and pos2word2 gram shows strong level of agreement. Figure 1 shows the Accuracy plot when comparing with different features with SVM. We also compared our results when SVM classifier used with other systems when naive bayes and random forest methods applied as shown in Table 2.

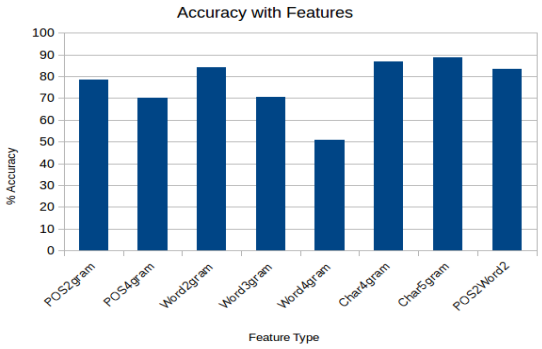

Figure 1. Accuracy plot when comparing with different features with SVM

Table 2. Comparative result when using different classification methods

Feature Type	Accuracy		
	SVM	Naive Bayes	Random Forest
POS2gram	78.65	66.08	75.43
POS4gram	70.17	66.95	64.32
Word2gram	84.21	78.36	74.12
Word3gram	70.46	72.22	70.32
Word4gram	50.87	59.50	64.03
Char4gram	86.98	80.55	82.16
Char5gram	88.59	80.55	80.17

Across all the features type SVM does best. Only in case of word 4 gram, SVM shows poor result while naive bayes and random forest did good. When with naive bayes and random forest classification methods compared on big period dataset then random forest method did best for POS 2-gram and character 4-gram and for all others naive bayes did best.

While, the accuracy compared with the methods in [13] and [17], shown in Table 3. These methods are similarity based, and SVM is a machine learning algorithm. SVM outperforms the other two. In our dataset performance of SCAP is worst as compared to Feature sampling [13] and SVM.

Table 3. Comparing the results with the attribution method for character 4-gram

Methods	SCAP [17]	Feature Sampling [13]	SVM
Accuracy	57.44	84.89	86.98
Precision	58.13	86.12	88.25
Recall	48.69	82.33	83.46

5. CONCLUSION

In this work, we studied and implemented a system which is capable to handle authorship attribution problem for text written by authors at different time frame. Experiment was made on English text corpus written by author at different time-frame. Text corpus from the newsgroup collection, letter collections are considered. While performing experiments focus is given only on feature set, how features extracted and represented and the selection procedure of features during model building. In the experiments, the SVM method gives very good result as compared to navie bayes and random forest algorithm for character n-gram, word n-gram and pos n-gram. Highest achieved result was 88.59% with character 5-gram features. Combined features of pos 2-gram and word 2-gram produces results up to 83.33% where individually they produced 78.65% and 84.21% respectively. Dataset collected over long time period average 30 years of time span. Our experiment is based on feature selection criteria. We can broaden our view to make our system more robust. In future work can be extended by categorizing different corpus type. Improving system by increasing accuracy of system.

REFERENCES

[1] Min Yang, Kam-Pui Chow, "Authorship attribution for forensic investigation with thousands of authors," *IFIP International Information Security Conference, Springer*, Berlin, Heidelberg, pp. 339-350, 2014.

[2] Tamboli M. S. and Prasad, R. S., "Authorship analysis and identification techniques: A review," *International Journal of Computer Applications*, vol. 77, no. 16, 2013.

[3] Rajesh Prasad, U. V. Kulkarni, and Jayashree R. Prasad, "A novel evolutionary connectionist text summarizer (ECTS)," *Anti-counterfeiting, Security and Identification in Communication, ASID 2009. 3rd International Conference on. IEEE*, 2009.

[4] Lancashire I. and Hirst G., "Vocabulary changes in Agatha Christie's mysteries as an indication of dementia: A case study," *19th Annual Rotman Research Institute Conference, Cognitive Aging: Research and Practice*, pp. 8-10, 2009.

[5] Can F. and Patton J. M., "Change of writing style with time," *Computers and the Humanities*, vol. 38, no. 1, pp. 61-82, 2004.

[6] Van Dam M. and Hauff C., "Large-scale author verification: temporal and topical influences," *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, ACM 2014, pp. 1039-1042, 2014.

[7] Stamatatos E., "A survey of modern authorship attribution methods," *Journal of the Association for Information Science and Technology*, vol. 60, no. 3, 538-556, 2009.

[8] Azarbonyad H., Dehghani M., Marx M. and Kamps J., "Time-aware authorship attribution for short text streams," *In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 727-730, 2015.

[9] Zhai C. and Lafferty J., "A study of smoothing methods for language models applied to ad hoc information retrieval," *In ACM SIGIR Forum, ACM 2017*, vol. 51, no. 2, pp. 268-276, 2017.

[10] Keikha M., Gerani S. and Crestani F., "Time-based relevance models," *In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, ACM 2011*, pp. 1087-1088, 2011.

[11] Zheng R., Li J., Chen H. and Huang Z., "A framework for authorship identification of online messages: Writing-style features and classification techniques," *Journal of the Association for Information Science and Technology*, vol. 57, no. 3, pp. 378-393, 2006.

[12] Abbasi A. and Chen H., "Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace," *ACM Transactions on Information Systems (TOIS)*, vol. 26, no. 2, pp. 7, 2008.

[13] Koppel M., Schler J. and Argamon S., "Authorship attribution in the wild," *Language Resources and Evaluation*, vol. 45, no. 1, pp. 83-94, 2011.

[14] Ding S. H., Fung B. and Debbabi M., "A visualizable evidence-driven approach for authorship attribution," *ACM Transactions on Information and System Security (TISSEC)*, vol. 17, no. 3, pp. 12, 2015.

[15] Kestemont M., "Function Words in Authorship Attribution. From Black Magic to Theory?," In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pp. 59-66, 2014.

[16] Stamatatos E., Fakotakis N. and Kokkinakis G., "Automatic text categorization in terms of genre and author," *Computational linguistics*, vol. 26, no. 4, pp. 471-495, 2000.

[17] Frantzeskou G., Stamatatos E., Gritzalis S., Chaski C. E. and Howald B. S., "Identifying authorship by byte-level n-grams: The source code author profile (SCAP) method," *International Journal of Digital Evidence*, vol. 6, no. 1, pp. 1-18, 2007.

[18] Kešelj V., Peng F., Cercone N. and Thomas C., "N-gram-based author profiles for authorship attribution," In *Proceedings of the conference pacific association for computational linguistics*, PACLING 2003, vol. 3, pp. 255-264, 2013.

[19] Kocher M. and Savoy J., "A simple and efficient algorithm for authorship verification," *Journal of the Association for Information Science and Technology*, vol. 68, no. 1, pp. 259-269, 2017.

[20] Burns K., "Bayesian inference in disputed authorship: A case study of cognitive errors and a new system for decision support," *Information Sciences*, vol. 176, no.11, pp. 1570-1589, 2006.

[21] Cheng N., Chandramouli R. and Subbalakshmi K. P., "Author gender identification from text," *Digital Investigation*, vol. 8, no. 1, pp. 78-88, 2011.
[22] Argamon S., Šaric M. and Stein S. S., "Style mining of electronic messages for multiple authorship discrimination: first results," *In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM 2003*, pp. 475-480, 2003.
[23] Rocha A., Scheirer W. J., Forstall C. W., Cavalcante T., Theophilo A., Shen B. and Stamatatos E., "Authorship attribution for social media forensics," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 1, pp. 5-33, 2017.
[24] Tamboli, Mubin Shoukat and Rajesh S. Prasad., "Feature Selection in Time Aware Authorship Attribution" *International Conference on Advances in Communication and Computing Technology (ICACCT)*. *IEEE*, 2018.
[25] Mohammed AL-Sarem, Abdel-Hamid Emara, "The effect of training set size in authorship attribution: application on short Arabic texts," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 1, pp. 652-659, 2019
[26] Kestemont M., "Function Words in Authorship Attribution. From Black Magic to Theory?," In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pp. 59-66, 2014.
[27] Sunil Kale and Rajesh S. Prasad, "Author Identification using Sequential Minimal Optimization with rule-based Decision Tree on Indian Literature in Marathi," *Procedia Computer Science*, vol. 132, pp. 1086-1101, 2018.
[28] M. Rizzo Irfan, M. Ali Fauzi, Tibyani, Nurul Dyah Mentari, "Twitter Sentiment Analysis on 2013 Curriculum Using Ensemble Features and K-Nearest Neighbor," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 6, pp. 5409-14, 2018
[29] Diederich Joachim, et al. "Authorship attribution with support vector machines," *Applied intelligence*, vol. 19, no. 1-2, pp. 109-123, 2003.