RESEARCH ARTICLE OPEN ACCESS

A Brief Analysis of Authorship Attribution

Urmila Mahor *, Aarti Kumar **
**(Department of Computer Science and Engineering, Rabindra Nath Tagore University, Bhopal)*
*** (Department of Computer Science and Engineering, Rabindra Nath Tagore University, Bhopal)*

ABSTRACT

To identify the ownership of a document by its style of writing is a cumbersome task. In the digital era, it is very easy to copy someone's document and publish it with its name. So it is very necessary to identify the fake authorship. Authorship attribution becomes difficult when it is performed manually. However, this process needs automation, when document size becomes large. The primary goal of our study is to analyze the important role of various features, Attribution techniques, and corpus size based on the text document written by different authors in authorship attribution. This paper focuses on the various features, methods, and impact of Corpus size in authorship attribution. In this paper, we mentioned the types of various features, methods of feature selection, and classification methods for authorship attribution (AA). We learned of the importance of data or corpus size, which plays a heuristic role in the choosing of a sufficient amount of features to appropriately recognize the true writer of an undesignated text or piece of document. We also analyzed that by the time features type also changed and new features were introduced that made this task easier

***Keywords* -** Authorship attribution, corpus size, features, classification.

Date of Submission: 17-07-2022 Date of Acceptance: 01-08-2022

I. INTRODUCTION

Every writer has his or her writing habits, which makes them different from one another. Content of documents depends on topic, genre, and area. besides these, each author has different unique writing characteristics known as stylometry. Stylometry means the art of writing, using sentence length, word length, special quotes and lines, use of poetry, use of examples, active or passive voice, different types of speeches like direct or indirect speech, some special types of symbols, use of question words and emotional phrases, etc. Each person's writing style depends on mood, topic, and language. A person can express his/her idea using different words or in particular words. When an author writes content anonymously then he/she uses some words frequently/habitually unconsciously. When two or more authors claim the authorship for the same piece of document, in this case, AA helps to know the genuine ownership of content by identifying the author's writing style or technique. AA is the latest field of forensics nowadays. Authorship attribution is a combination of art and science that helps to find the genuine author of an unknown text/document, based on its stylistic features. These features can judge the author's gender, age, ideology, and religion, or motivation, mood, education. There are several types of stylistic features such as character, lexical, Syntactic, Structural, and Semantic. Most researchers used various combinations of the features for the authorship attribution task

II. INDENTATIONS AND EQUATIONS

Authorship attribution has been started in the late 18th century; First word was done on Shakespeare documents. Mosteller et.al. contributed considerably to the attribution of "The Federalist Papers‟. The statistical language modeling techniques of Mosteller et.al. were the first work that published in this field. Authorship attribution started in eighteen century and various researchers used different parameters for judging the authorship attribution. Initially, word & sentence length were used as features. Later on researcher used word frequency count, character frequencies, function words, vocabulary richness, and graph-based methods as measures for studying authorship attribution. Mingzhe et al. emphasized the use of a comma (,) as a feature, as it is used by an author as a breakpoint in a sentence to clarify pause in a sentence and is being used differently by different authors.

Andrew et al. used novel topic cross-validation for measuring the authorship in their work. Cross-validation is performed on the unknown topics of the training data set. Precision and recall F-measures were used for finding the results. Ali Osman Kausakci proposed the k-NNRV method as a new tool to deal with the variations in the writing. This method helps in recognizing the new features.

Esteban et al. focused on phrase-level lexical-syntactic features and graph-based representation of lexical features of word such as prefixes, suffixes, and stop words. Character features like vowel combinations and permutations were also used. They found that graph-based representation performed better than other ones.

Agramon et al. proposed a new feature of systemic functional linguistics(SFL) to analyze the text. In this, they used the frequencies of conjunction, modality, and comment. Systemic Functional Linguistics (SFL) provides a base for stylistics feature selection.

Michael Gamon used shallow linguistic analysis and deep linguistic analysis features and concluded that deep linguistic analysis features in authorship attribution reduced the error rate over the function word frequencies.

III. FEATURES TYPES

There are generally five types of features

1. Lexical Features 2. Character Features 3.Syntactic Features 4. Structural Features 5. Content Based Features

3.1 Lexical features - The first proposed works in authorship attribution had been focused on word length, syllables per word, and sentence length. For these one needs a tokenizer, stemmer, and lemmatizer to handle such types of features. In 1887 Mendenhall proposed new measures like average word length for authorship work. Later on, Mendenhall's work was followed by Yule and Morton. they used sentence length as a feature for authorship identification. Lexical features are based on the word, sentence, or paragraph-length count. A bag of words is also an approach particularly used for the selection of lexical-based feature sets. According to N. Akiva and M. Koppel there are two main trends in lexically-based approaches: 1) Those that represent the vocabulary richness of the author and 2) those that are based on number of occurrence of individual words. The selection of the specific function words to use as features is generally based on some criteria. Various sets of function words have been proposed for English Abbasi and Chen proposed a set of 301 features, a set of 303 feature words were used by Argamon, Saric, and Stein in their work, and Zhao and Zobel used a set of 363 function words, Koppel and Schler proposed a set of 480 function words, another set of 675 words were used by Argamon, Whitelaw, Chase, Hota, Garg, and Levitan.

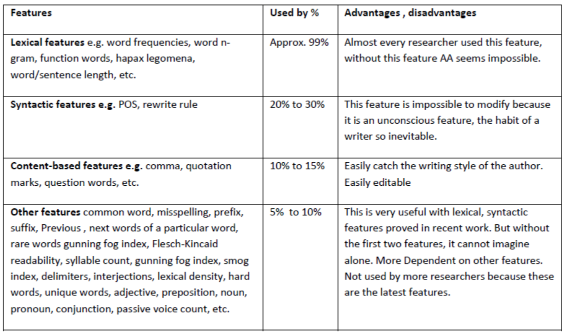
3.2. Character Features- The character features have also been used for the authorship attribution. This focus is on those characters which are frequently used by an author in his/her work such as quotation marks, apostrophes, commas, semicolons, upper case, and lowercase characters and punctuation marks. They are counted based on a per sentence or per paragraph basis. These are normally used along with lexical or syntactic feature-based methods. Kjell used character n-gram feature selection with the nearest neighbor and Naïve Bayes classifier.

3.3. Syntactic Features - Syntactic features are related to the formation of a sentence or the structure of the sentence. For this, we require a parser, sentence splitter, or chunker to represent a sentence structure. Some researchers have shown that the use of lexical features and syntactic features together improves the performance of authorship attribution as compared to individual ones. Syntactic features are noun, verb, length of the noun, length of verb phrases counts, etc. Koppel and Schler proposed the use of the syntactic features in 2003 based on syntactic errors such as sentence fragmentation, run-on sentences, mismatched tense, etc. Karlgren and Eriksson focused on model-based features such as syntactic features or adverb expression and presences of clauses in the sentences for authorship attribution.

3.4. Structural Features- Structural features represent the organization of content and its structure. This feature demonstrates deeply the writes content organization like paragraph margin, paragraph length, difficulty levels of reading and writing, presence of images, hyperlinks, beginning and ending salutation, and use of commas and apostrophes frequently and non-frequently.

3.5. Content Specific Features- In a Particular domain topic, a specific set of words will come regularly those words are called Content-specific features. While discussing diseases some words like treatment, therapy, and medicines will appear these words are treated as content-specific features. The semantic content of a document is less effective because they are variable-dependent and easily editable and under the conscious control of the author. While semantic features are difficult to manipulate so they are more useful as compare to content features.

Table 1 : Show the percentage of features used by the authors in their work



Features	**Used by %**	**Advantages , disadvantages**
Lexical features e.g. word frequencies, word n-gram, function words, hapax legomena, word/sentence length, etc.	Approx. 99%	Almost every researcher used this feature, without this feature AA seems impossible.
Syntactic features e.g. POS, rewrite rule	20% to 30%	This feature is impossible to modify because it is an unconscious feature, the habit of a writer so inevitable.
Content-based features e.g. comma, quotation marks, question words, etc.	10% to 15%	Easily catch the writing style of the author. Easily editable
Other features common word, misspelling, prefix, suffix, Previous , next words of a particular word, rare words gunning fog index, Flesch-Kincaid readability, syllable count, gunning fog index, smog index, delimiters, interjections, lexical density, hard words, unique words, adjective, preposition, noun, pronoun, conjunction, passive voice count, etc.	5% to 10%	This is very useful with lexical, syntactic features proved in recent work. But without the first two features, it cannot imagine alone. More Dependent on other features. Not used by more researchers because these are the latest features.

IV. CORPUS SIZE

Dataset Characteristics- As we studied various research papers and we found that corpus size also matters in authorship attribution, Higher the corpus size, the higher will be the accuracy of the classification result. Till now who has not told us about the perfect corpus size and how much data should be there so that we can get the right accuracy. According to the study, 10,000 words per author has traditionally been considered a reliable minimum fund size for an authoritative work. Based on a survey on text size, it was found that 5,000 words were considered the minimum requirement for training purposes (Kim Luks and Walter Delman, 2010), and 200 words per author are considered for short text (S.. Argamon et al.) To fully evaluate an authorship attribution method, performance must be measured under a variety of conditions. Researchers use more than 10,000 words per author, which is considered a reliable minimum size for attribution (F. Howdy et al., Feiguin, Hurst, Halteren), and Schwartz et al. focused on the use of smaller text sizes. , like 100 to 500 words per document. As we studied the data size. There is a limit to the data. If the data is limited then attribution becomes difficult because insufficient information is not able to identify authorship. In this case, traditional approaches are less reliable. Short texts require reliable representations and machine learning algorithms that can handle limited data. Reducing the length of training samples has a direct impact on performance. It is very difficult to predict or declare text of any particular length to properly quantify the stylistic properties of anonymous text.

V. FEATURE SELECTION

5.1 Chi-square-based feature selection: χ2 is a popular feature selection algorithm. This term and the occurrence of the class are the two events [15]. Rank is assigned to each term according to

$$X^2(\mathbb{D},t,c) = \sum_{e_t\in\{0,1\}} \sum_{e_c\in\{0,1\}} \frac{(N_{e_te_c} - E_{e_te_c})^2}{E_{e_te_c}}$$

Where et =1 if the document contains term t and et =0 otherwise C is a random variable that takes values ec =1 if the document is in class C and ec =0 otherwise.

5.2 Correlation coefficient feature selection: The correlation coefficient is another approach for a pair of variables (X, Y), the linear correlation coefficient r is given by

$$r = \frac{\sum_i (xi - \overline{xi})(yi - \overline{yi})}{\sqrt{\sum_i (xi - \bar{xi})^2}\sqrt{\sum_i (yi - \overline{yi})^2}}$$

Where xi is the mean of X, and yi is the mean of Y .The value of r lies between -1 and 1, inclusive. If X and Y are completely correlated, r takes the value of 1 or -1, if X and Y are totally independent, r is zero.

5.3 Principal Component Analysis (PCA): The main reason for using principal components analysis is to derive new variables that are linear combinations of the original variables. Savoy and Jacques used PCA to distinguish the similarity and dissimilarity between the texts in computational terms. Researchers used PCA to resolve several outstanding authorship problems.

5.4 Latent Semantic Analysis: It is a well-known method for extracting the dominant features from large data sets and for reducing the dimensionality of the data. This corpus can be represented as a term-document matrix, which is obtained by constructing the new reductive feature space. Each document is represented by

$$d' = dTUk$$

Where d' is the new reduced feature vector and dT is the feature vector applied by the above-mentioned feature selection method.

5.5 SVM Feature Evaluator: Support Vector Machine (SVM) is well known for categorizing text. Zhang et al. used

SVM for authorship attribution. SVM is successful because of its good properties of regularization, maximum margin, robustness, etc.

VI. LEARNING AND EVALUATION

The performance of the different methods that are studied in this research is compared by calculating precision, recall, and F-measure.

Precision: Precision is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. In term of true positive and false positive, it is defined as

$$\text{Precision}=\frac{TP}{TP+FP}$$

Recall: Recall is the ratio of the number of relevant records retrieved to the total number of relevant records in the database. In term of true positive and false positive it is defined as

$$\text{Recall}=\frac{TP}{TP+FN}$$

F-measure is the harmonic mean of precision and recall. F-measure focuses on the positive class, even if inverted, are devalued compared to positive features. It is defined as

$$F=\frac{2Xprecision\ X\ Recall}{(Recall+precision)}$$

VII. CLASSIFICATION TECHNIQUES

7.1. Naive Bayes classifier

The Naive Bayes Classifier technique is a Bayesian theorem and is particularly suited when the dimensionality of the input is high. Despite its simplicity, Naive Bayes is often a sophisticated classification method. The Naive Bayes classifier builds a probabilistic model of each author class based on the training data for that class. Then it calculates and multiplies the probabilities of all the features to give the probability of the test text. The most likely of all authors is the author of that anonymous or test text. Generally, it has been observed that the Naive Bayes classifier has been used for attribution of authorship in many languages including English [Hoorn et al, 1999; Zhao and Zobel, 2005]. The disadvantage of Naive Bayes is when the test data contains features that the model has not observed in the training data. So some probabilities give zero results because none of the training data falls into the range. These null counts have zero probability, making the Naive Bayes classifier unable to predict a class.

7.2. Sparse classifier

Sparse Representation Classification (SRC) is a classical method. It was proposed in 2009. This method has been successful in the face recognition task. Many researchers adopted this method for work in authorship verification work. The method consists in identifying the components of an unauthorized document from the authors' multiple documents. It also has some disadvantages. The classification criterion of SRC is from residues using the truncation method. If the ith residual is the smallest, then the SRC judges that the test sample belongs to the ith class. However, by way of truncating, some test samples are also misclassified by SRC.

7.3. Nearest Neighbor classifier

The nearest neighbor classifier achieves consistently high performance, without preconceptions about the distribution from which the training examples are drawn. It includes a training set of both positive and negative cases. The nearest neighbor algorithm does not build a model from the training data. So, this algorithm is called 'lazy learner'. It obtains information from test data when it is necessary to classify an unseen sample. Each sample from the training data is represented as an n-dimensional data point. 'N' represents the number of features used to describe the data. When an unseen sample is introduced into the algorithm it will retrieve the k-nearest neighbors of that sample which is computed with a proximity measure. `k' is the number of nearest neighbors that should be retrieved. The unseen data sample gets the same class label as its k neighbors. If these neighbors have more than one class then the unseen data gets the label that most of its neighbors have. If there is a tie between class labels, the unseen sample is assigned a random class label. One disadvantage of the nearest neighbor algorithm is that when there are too many features Hoorn et al. Comparison of a neural network with Bayes classification and the nearest neighbor algorithm in identifying the author of a poem. KNN [Zhao and Zobel, 2005; Luyckx and Daelemans, 2008].

7.4. Neural Networks classifier

A neural network is composed of nodes with directed weighted links between them. The network consists of an input layer representing the input features, an output layer to output the model, and possibly several hidden layers. The weighted sum of the inputs of a node is used as the input to an activation function, which determines the output of that node. The activation function makes it possible to produce an output that is a non-linear function of the input. During the learning phase, the weights of the network are adjusted until the error rate is minimized. A widely used method to reduce this error is gradient descent. A commonly used method for training hidden units is back-propagation. Kjell et al. used a neural network with the character bigram to identify authors of articles in Federalist papers. The disadvantage of a neural network is to set too many parameters, the number of input nodes that depend on the number and type of features, the number of output nodes that depend on the number of classes, the number of hidden layers, and the number of activation functions. And the initial weights are hidden in the nodes in the layers. Improper setting of these parameters can result in under-fitting, so the network cannot fully describe the data or over-fitting, so the network cannot generalize well to unseen data.

7.5. Support Vector Machines classifier

This technique is based on finding the maximum margin hyperplane that separates the data into two sets. Finding this hyper-plane is based on structural risk minimization, a principle that seeks to minimize generalization error while minimizing training error and avoiding very complex models. The machine learning techniques discussed earlier only reduce the training error, but this does not mean that the generalization error is minimized. So theoretically this means that SVM can perform better generalizations on unseen data [Argamon, 2008]. The standard authorship attribution in which we are required to assign an anonymous document to one of a small closed set of candidates is well understood and has been summarized in several surveys [Stamatos, 2009]. A binary learning problem and SVM have often been found to perform well for binary authorship problems [Abbasi and Chen, 2008; Zhang et al., 2006] SVM [Diederich et al., 2000; Gamon 2004].

7.6. LDA classifier - Latent Dirichlet allocation (LDA) [Bleigh et al., 2003] is used to model authors from their texts. LDA is a generative probabilistic model traditionally used to find topics in textual data. The main idea behind LDA is that each document in the corpus is generated from a distribution of topics, and each word in the document is generated according to a per-topic word distribution. [Bleigh et al., 2003] showed that the use of LDA can improve the performance of supervised text classification for dimensionality reduction. We know of only one case where LDA was used in authorship attribution [Prince et al., 2009] reported preliminary results on using LDA subject distributions as feature vectors for SVMs, but He did not compare the results obtained with LDA-based SVMs with the results obtained. with SVMs trained directly on tokens [Yanir Cersei et al., 2011].

In LDA, the generation of a document collection is modelled as a three-step process.

First, for each document, a distribution over topics is sampled from a Dirichlet distribution.

Second, for each word in the document, a single topic is chosen according to this distribution.

• Finally, each word is sampled from a polynomial distribution of words specific to the sample topic [Michael Rosen-Zwi et a., 2011].

In LDA, the generation of a document collection is modeled as a three-step process.

• First, for each document, a distribution over topics is sampled from a Dirichlet distribution.

• Second, for each word in the document, a single topic is chosen according to this distribution.

• Finally, each word is sampled from a polynomial distribution on words specific to the sample topic [Michael Rosen-Zwi et a., 2011].

7.7 Decision Trees classifier - These are simple but successful inductive learning methods. In a decision tree, the features of the data are modeled as a tree structure. The root node contains a feature test that isolates data samples that have a different value for the feature being tested. Each test should result in a subset of possible categories. The number of decision trees that can be built is exponential in the number of features. Therefore an algorithm building decision trees needs to use a strategy that produces a tree within a reasonable amount of time. A commonly used strategy is a greedy approach, which locally builds the nodes of the decision tree by choosing the most optimal test. There are several ways to decide what the most optimal test is. Possible measures are 'Gini index' and 'classification error'. One advantage of decision trees is that once the tree is built, the classification of unseen data is much faster. Another advantage is that when one is chosen as a test when two features are highly correlated, the other will not be used. One disadvantage of decision trees is that they can be used in a decision tree when the data contains irrelevant features, resulting in a result larger than the three required for classification [Zhao and Zobel, 2005].

VIII. CONCLUSIONS

In this paper, we studied various types of features and feature selection methods, various types of classification methods, and their uses. If data is linear then which classifier should be applicable and will provide good results and when data is nonlinear then which will be applicable. We also studied the size of data or corpus, because if less corpus size is not capable enough to predict the accurate result, if data is small size then traditional approaches are not helpful. We also studies various evaluation measures that are generally used in finding the accuracy of training the classifier, According to our study we found that the most used features are lexical and syntactic features, We cannot perform the Authorship attribution with only one type of feature, we need the combination of two or three types of features to justify the authorship.

REFERENCES

[1]. N. Akiva , M. Koppel , A generic unsupervised method for decomposing multi-author documents, *Journal of the American Society for Information Science and Technology*, 64(11) , 2013,2256–2264.

[2]. K. Aldebei, X. He, and J. Yang, Unsupervised decomposition of a multi-author document based on naive bayesian model, ACL, volume 2, Short Papers, ,2015, 501.

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

[4]. A. Abbasi and H.Chen, Applying Authorship Analysis to Extremist Group Web Forum Messages, Published by the IEEE Computer Society, 2005.

[5]. A. Abbasi and H. Chen , Visualizing Authorship for Identification, S. Mehrotra et al. (Eds.): ISI 2006, LNCS 3975, © Springer-Verlag Berlin Heidelberg, 2006)

[6]. [6] S. Argamon, M. Saric, and S.Stein, Style mining of electronic messages for multiple authorship discrimination, *First results, Proc. of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003.

[7]. S. Argamon, C. Whitelaw, and P. Chase, Stylistic text classification using functional lexical features, *Journal of the American Society of Information Science and Technology*, 58(6),2007, 802–22.

[8]. S. Argamon and Levitan, Measuring The Usefulness Of Function Words For Authorship Attribution, *Proc. of the ACH/ALLC Conf.*, 2005.

[9]. S .Argamon, M. Koppel, Pennebaker, Schler , Automatically Profiling the Author of an Anonymous Text, Communications of the ACM 52 (2) , 2013, 119-123.

[10]. J. Hoorn, S.Frank, W. Kowalczyk, and F. van der Ham, Neural network identification of poets using letter sequences, Literary and Linguistic Computing, 14(3) , 1999, 311-338.

[11]. W. Daelemans, and A. van den Bosch, Memory-Based Language Processing, Studies in Natural Language Processing, Cambridge University Press, 2005.

[12]. J. Diederich, J. Kindermann , E. Leopold, and G. Paass, Authorship attribution with support vector machines, Applied Intelligence, 19(1–2),2000, 109–23.

[13]. B. Kjel, W. A. Woods, O. Frieder, Information retrieval using letter tuples with neural network and nearest neighbor classifiers, *In IEEE International Conf. on Systems*, Man and Cybernetics, Vancouver, BC, 1995, vol. 2, 1222-1225,.

[14]. F. Howedi and M. Mohd, Text Classification for Authorship Attribution Using Naive Bayes Classifier with Limited Training Data, Computer Engineering and Intelligent Systems ISSN 2222-1719, Paper ISSN 2222-2863, vol.5, No.4, 2014.

[15]. O. Feiguin and G. Hirst, Authorship attribution for small texts: Literary and forensic experiments, *Proc. of the SIGIR International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection*, PAN 2007, Amsterdam, Netherlands, July 27, 2007

[16]. M. Gamon, Linguistic correlates of style: Authorship classification with deep linguistic analysis features, *Proc. of the International Conference on Computational Linguistics*, 2004,.611–617.

[17]. G. Hirst, and O. Feiguin, Bigrams of syntactic labels for authorship discrimination of short texts, Literary and Linguistic Computing, 22(4),2007, 405–17.

[18]. M. Koppel, J. Schler, Authorship verification as a one-class classification, 54 Literary and Linguistic Computing, Machinery, 2004, 489–95.

[19]. M. Koppel, J. Schler, S. Argamon, Y. Winter, The “Fundamental Problem” of Authorship Attribution, Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK, 2012.

[20]. M. Koppel, J. Schler, S. Argamon, and E.Messeri, Authorship attribution with thousands of candidate authors, *Proc. of the 29th Annual International ACM SIGIR Conference on Research and Development on Information Retrieval. Seattle*, WA: Association for Computing Machinery,2006 659–60.

[21]. M. Koppel, Computational Methods in Authorship Attribution, JASIST, 60 (1), 2008, 9-26.

[22]. M. Koppel, J. Schler, Exploiting Stylistic Idiosyncrasies for Authorship Attribution, *Proc. of. IJCAI, Workshop on Computational Approaches to Style*, 2003.

[23]. M. Koppel, J. Schler, D. Mughaz, Text Categorization for Authorship Verification, *Twenty-first International Conference on Machine Learning*,2004

[24]. K. Luyckx, , W. Daelemans, Authorship attribution and verification with many authors and limited data, *Proc. of the 22nd International Conference on Computational Linguistics – vol. 1, ser.COLING, Stroudsburg, PA, USA: Association for Computational Linguistics*, 2008,513–520.

[25]. K. Luyckx, , W. Daelemans, The effect of the author set size and data size in authorship attribution, *Journal, Literary and Linguistic Computing*, vol. 26., Issue 1, Pagination, ISSN, 0268-1145, 2012,35-55.

[26]. E. Stamatatos, A survey of modern authorship attribution methods, *Journal of the American Society for Information Science and Technology*, 60(3),2009, 538–56.

[27]. E. Stamatatos, N. Fakotakis,n and G. Kokkinakis, Text genre detection using common word frequencies, *Proc. of the 18th International Conference on Computational Linguistics*, vol. 2.Saarbrucken, Germany: Association for Computational Linguistics, 2000, 808–14.

[28]. E. Stamatatos, Fakotakis, Kokkinakis, Automatic authorship attribution, *Proc. of EACL*, 1999.

[29]. [29] E. Stamatatos, On The Robustness Of Authorship Attribution Based On Character N-Gram Features, Am. Soc. Inf. Sci. Technol. 60,2009, 538–556.

[30]. E. Stamatatos, Authorship Attribution Based On Feature Set Sub spacing Ensembles, *International Journal on Artificial Intelligence Tools*, vol. 1010, No. 10, 2006, 1–16.

[31]. W. B. Cavnar and J. M. Trenkle, N-Gram-Based Text Categorization, *Journal of Statistical Software*, 52 (6), ISSN 1548-7660, 1994,1-17.

[32]. M. Zhang, J. Yao, A rough sets based approach to feature selection, *Proc. 23rd International Conf. of NAFIPS*, 2004, 434–439.

[33]. Y. Zhao, J. Zobel, Effective and scalable authorship attribution using function words, *Proc. of the 2nd Asia Information Retrieval Symposium. Jeju Island, Korea: Springer*, 2005, 174–90.

[34]. Y. Zhao, J. Zobel, Searching with Style: Authorship Attribution in Classic Literature, *Proc. of 30th Australasian Conference on Computer Science*, vol. 62, 2007, 59–68.

[35]. Y. Zhao, Effective Authorship Attribution in Large Document Collections, *Proc. of the 16th international conference*, ACM. doi:10.1145/1242572. 1242659, 2007, 639-648.

Urmila Mahor, et. al. “ A Brief Analysis of Authorship Attribution.” *International Journal of Engineering Research and Applications (IJERA*), vol.12 (07), 2022, pp 186-193.