AA with Tf-Idf weight using ML

Urmila Mahor*, Aarti Kumar**

* Computer Science and Engineering Department, Rabindranath Tagore University
** Computer Science and Engineering Department, Rabindranath Tagore University

Abstract- Authorship attribution is the process of automatically identifying a document's author by analysing their writing style. Its history is extensive, and there are many different applications for it. As online work increases, it plays a vital role in forensic science, the detection of plagiarism, conflicts between authors and research. Very helpful when two people contend ownership of the same material. This is a classification kind of problem. It is not in line with the goal of text categorization because it simply takes into account the author's erratic writing style, regardless of whether it uses text categorization techniques for text pre-processing. It is highly dependent on the writing style characteristics of the authors for the author attribution task to be successful. In order to determine a writer's writing technique, multiple researchers have suggested different kinds of characteristics, including language, syntax, semantics, and content-based features. Several of these features have been applied to categorize articles. In this study, Support vector machines, parametric and nonparametric techniques, supervised and unsupervised techniques, and TF and IDF with FW and stylometric characteristics were also used. With the English corpora, we ran a variety of experiments. We conducted numerous tests with various feature sets retrieved from the corpus using various classifiers, and we then enhanced our success rate by integrating these outcomes. Based on the feature sets we evaluated, we identified the classifiers that produce reliable results. Experimentally, success rates change dramatically when feature sets are combined. However, the models that are tested by support vector classifiers (SVC) with a BoWs, FWs, and Gaussian.

Index Terms- Sylometric Features, TW, BoWs, FWs, classification, Authorship Attribution.

I. INTRODUCTION

A method for identifying the real author of an anonymous article from a list of possible candidates is known as authorship attribution. We can deduce the original ownership of a given unnamed document by investigating documents from a given set of documents by authors. Nowadays authorship attribution has become very popular and useful in many fields like security, forensic analysis, plagiarism research, and plagiarism detection. Researchers are copying and claiming others' content in their names since digitalization research articles are available online. Authorship attribution has two types: open and closed group. The vast majority of published research in this field focuses on closed group set,where each possible author is predetermined and known [10,16]. A writer can be recognized by the way he writes using stylistic analysis [10,17]. Some stylometric variables, like FWs (Ex preposition(s), article(s), etc.) and lexical characteristics are not subject-shift resistant and unaffected by topic and genre [11,14]. As we researched [10,15,18], character n-grams showed to be another trait that is very useful for authorship attribution. According to [10,19] syntactic elements are especially crucial since authors employ them unconsciously and cannot readily change them in their writing (which is why they are topic- and genre-independent features). We discovered that the majority of authorship attribution research concentrates on how well each style-based feature performs. In this study, we show that the model performs better when style-oriented features are combined. In addition, combining many feature sets yields superior outcomes versus employing a single feature type. Our research in this study is focus on the closed-set authorship challenge, In this model is trained with known candidate authors and the aim is to select the correct candidate author from among a given group of candidates for an unknown piece of documentation using stylometry.

II. RELATED WORK

In the 1980s, work on authorship attribution started. In 1887, Mendenhall put forward the first piece with a feature relating to word length. The Federalist Papers were written in 1964, in this Mosteller and Wallace employed Bayesian Statistical Analysis (BSA) with FWs to determine who wrote them. Nowadays, authorship attribution plays a significant role in a wide variety of applications like, forensics, legal matters, plagiarism checking, and investigations. Authorship attribution research has grown significantly in recent years, incorporating techniques like Natural- Language Processing (NLP), Machine Learning (ML), text mining, deep learning (DL), Neural Networks (NNs), Deep Belief Networks (DBNs), information retrieval, as well as linguistic properties like character, word, and sentence levels. The study of style and authorship attribution in distinctive and uplifting works is termed stylometry. Average sentence length, average paragraph length, question words, direct-indirect speech usage, active and passive voice, grammatical faults, idiom usage, and many other stylistic characteristics are examples of stylometric traits. In their work, L. Tanguy et al. utilized decision trees and rule-based learners, maximum entropy, and word-level and sentence-level data. In terms of correctness, they found that the maximum entropy technique outperformed ML algorithms[1,19]. Prefixes, suffixes, word lengths, and BoWs

were all utilized by the researchers [2, 3,13]. have been demonstrated
in numerous studies to be useful authorship traits [4,5,6,12].

III. OUR APPROACH

According to numerous researches, BoWs (Bag of Words) and FWs have typically been employed as document vectors. The TF-IDF is used alongside BoWs and FWs in our work. In order to do this, we extract common terms from the corpus that can be utilised to distinguish the authors' writing styles using pre-processing methods like tokenization and stemming. After that, we generate a document vector for each author independently. A candidate group of well-known authors is offered for the experiment work ([7][8]). The purpose of this work is to narrow down the closed set of documents from candidate writers to determine the true authorship of an unidentified text.

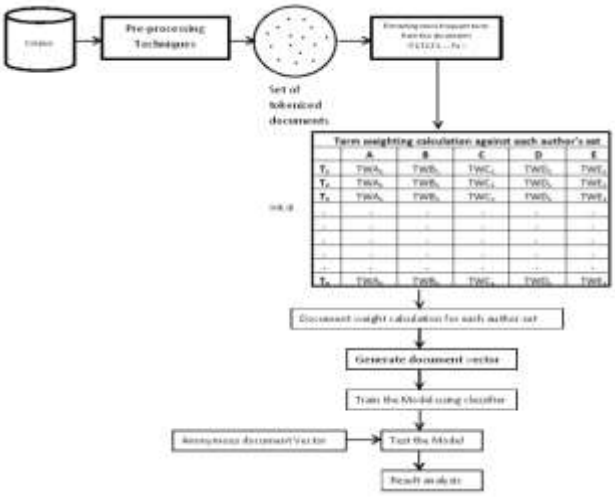

Figure. 1 Proposed model for work

A. Bag of Words(BOWs)

The frequent prevalent words in the corpus are referred to as attribute and are referred to as the "Bag of Words". Every document in the data set is represented in this form. The classification model is trained using document vectors, where each value represents the term's weight.

B. Function Words(FWs)

The relative importance of FWs in authorship attribution (particle, article,pronoun, conjunction).

Researchers	Year	Success rate	Classifier
S. Argamon and Levitan	2000	90%	Support Vector Machine(SVM)
Stamatatos, Fakotakis, Kokkinakis	2001	65%-72%	Multiple Regression(MR), Discriminant Analysis (DA)
J. Diederich	2003	60%-80%	SVM
Kjell	1994	80-90%	Neural Networks (NNs), Bayesian Classifier (BC)

[8]

C. Vocabulary Level Classifier

By determining the degree of dissimilarity between the texts, the cosine similarity is utilized to distinguish between them. The cosine similarity formula determines how similar two vectors are to one another. The cosine angle between two vectors determines, if they point in the same direction or not. It is typically used in text processing to compare two texts' similarities. Assume that the cosine function is used to calculate the two vectors x and y. The degree to which the unattributed article variable resembles the writer's variables has been determined using the cosine relationship function.

$$\text{Similarity}(x,y)=(x.y)/(|x||.||y||)$$

D. Generation of Document Vector

A word vector, or WV, is used to represent every document. The term frequency (TF) and inverse document frequency (IDF) are combined to form an aggregate weighting for each word in each document. The weight of each term t in document d has been determined with the TF/IDF weight method using the following formula:

$$W_{(t,d)} = TF_{(t,d)} \times IDF_t$$

Where TF (t,d) is the term-frequency of t in d and is determined by:

$$TF_{(t,d)} = \frac{\text{number of times term(t) apperas in d}}{\text{total number of terms(words) in d}}$$

The formula for IDF(t), or inverse document frequency(IDF), is as follows:

N- All documents in the corpus

df- All documents that include word (t)

$$IDF(t) = \log \frac{N}{1+df}$$

Each term's TF and IDF weight is displayed as a term's weight wt. Using the word vector (WV) and the TF/IDF weights, determine the weight expression for the ith document (d) as follows:

$$w_{RP}(d_i) = \sum_t w_{t,d}\ \text{wv(t)}$$

D. Evaluation Measures

The researchers examine the Authorship Attribution efficacy (Accu) of author prediction using a variety of metrics, including recall, precision, F1 measure, and accuracy. Classification accuracy (Accu) measures, how often a classifier makes the right

choice. As it's opposite, the error rate (Err) displays the proportion of wrong choices.

$$Accu = (TP + TN)/(TP + TN + FP + FN)$$

$$Err = (FP + FN)/(TP + TN + FP + FN)$$

IV. PROPOSED PROCEDURE

Our proposed approach for this work

1 The source of our corpora is PANCLEF, which can be found on its official website. The legal corpus for this task is provided on this website.

2 first, we process the text using pre-processing methods on the corpus by following Stemming, parsing, removing stop words, collecting function words(FWs), Identifying phrases that appear frequently in the corpus (at least five times), collection of bigrams, unigrams, and rare words.

3 Calculate term weights for each set of writers in the texts.

4 By summing the weights for every word in document using document weight measure, the document weight for each group of authors is determined.

5 Use document weights to generate document vectors and train a classification model.

6 Apply testing

7 Analysis of the Results and perform comparison.

D1, D2,.....Dk indicates a group of papers here, T1, T2,.....Tl is a group of vocabulary terms, while (A, B, C, D, and E) is a group of five author categories. The term TL weights for author groups (A, B, C, D, and E) are TWAm, TWBm, TWCm, TWDm, TWEm, respectively.We also employed Hapex legomena,avg. judgment length, number of semicolons(;), ages(.), commas(,), question marks(?), andavg. paragraph length in a judgment among these trials. In tests, the classification model output from the confusion matrix is used to decide the actual writer of an unattributed text or document. Our key concern in this is the selection of suitable weight measures for calculations.

V. EXPERIMENTAL WORK AND RESULT ANALYSIS

In this work, Weka 3.7 tool was used for our experiment, and we used different techniques to complete our study, but we only needed four to get a successful result. We used following classifiers with three feature sets bag of words, function words and stylometric features in our work.

A. Naïve Bayes Classifier

This was applied with 4 distinct data fractions, and the result is good. The benefit of this approach is that both continuous and discrete data may be used with it, and it only needs a small quantity of training data.

Table 1. NB's results for 10 fold Cross

Success Rate with Data Ratio				
No. of Iteration	(90:10)	(80:20)	(70:30)	(60:40)
1	84.22	86.13	79.83	67.63
2	84.18	85.78	79.98	67.71
3	84.23	86.07	81.02	67.21
4	83.98	85.46	80.07	67.34
5	84.76	85.83	80.71	67.47
6	83.96	85.95	80.15	68.26
7	84.44	85.74	80.17	68.41
8	83.96	86.01	80.06	66.99
9	84.33	85.93	81.01	66.95
10	83.96	86.09	80.01	66.99
Avg	84.20	85.90	80.30	67.50

B. Decision Tree Classifier

With this classifier, a tree is formed in this fashion, and the benefit of this algorithm is its simplicity, which requires less data preparation, can handle problems with many outputs, and is suitable for both category and numerical data. The outcome is good.

Table 2. DT's results for 10 fold Cross validation

Success Rate with Data Ratio				
No. of Iteration	(90:10)	(80:20)	(70:30)	(60:40)
1	79.48	80.13	84.82	79.43
2	78.91	81	84.96	79.17
3	79.55	80.17	84.87	79.22
4	79.59	80.11	84.88	79.15
5	78.98	80.13	84.87	79.24
6	78.89	80.15	83.96	79.26
7	79.54	80.04	84.84	80.02
8	79.53	80.01	83.96	80.07
9	78.93	80.13	84.88	79.25
10	79.59	80.09	84.93	79.21
Avg	79.30	80.20	84.70	79.40

C. Support Vector Machine Classifier

Terms or words free of stop-words are extracted once the text has been pre-processed. We used the tf/idf method to calculate the weights for each phrase in this.

where,

tfk= The number of times the term "k" appeared in a document..
dfk = Number of documents containing the phrase "k"
D = A database's total quantity of records available [9].

Table 3: SVM's results for 10 fold Cross validation.

Success Rate with Data Ratio				
No. of Iteration	(90:10)	(80:20)	(70:30)	(60:40)
1	80.2	90.61	84.12	75.23
2	81.9	90.52	84.16	75.17
3	82.68	90.41	84.17	75.24
4	82.59	90.66	83.88	75.15
5	82.65	90.56	84.17	75.24
6	81.61	90.66	83.96	75.26
7	82.83	90.47	84.34	74.99
8	82.79	90.78	83.96	75.27
9	82.87	90.64	84.23	75.26
10	82.89	90.69	83.96	75.22
Avg	82.30	**90.6**	84.10	75.20

D. K-Nearest Neighbors(K-NNs) classifier

One of the most popular classifiers is a K-nearest neighbor, which is both a non-parametric method and conceptually straightforward. An unknown instance (I) is categorized in a training set according to the class of the nearest data point (k). To accomplish this, The distances among every record within the training sample and instance P are all calculated. Instance (I) is given the class that the majority of the nearest data points fall within for k>1. The "Manhattan Distance", "Euclidean Distance", and "Hamming Distance" are the most popular similarity measurements.

Table 4. KNNs iterative result for 10 fold Cross validation.

Success Rate with Data Ratio				
No. of Iteration	(90:10)	(80:20)	(70:30)	(60:40)
1	80.31	87.11	80.13	79.21
2	80.29	87.12	81.68	79.17
3	80.37	87.03	80.32	80.02
4	80.41	87.06	80.81	79.15
5	80.63	87.16	80.11	79.24
6	80.68	87.04	80.65	79.26
7	80.23	87.07	80.67	79.22
8	80.37	87.21	80.76	80.07
9	80.93	87.12	80.31	79.25
10	80.76	87.09	80.51	79.43
Avg	80.50	87.10	80.60	79.40

Following table 5 shows the combined result of all four classifiers with average success rate with different ratios of data.

Table 5. Success rate of different classifier

Success Rate of Model with classifiers				
Ratio of data	SVM Success Rate (%)	DT Success Rate (%)	KNNs Success Rate (%)	NB Success Rate (%)
90 -10	82.3%	79.3%	80.5%	**84.2%**
80-20	**90.6%**	80.2%	**87.1%**	**85.9%**
70-30	84.1%	**84.7%**	80.6%	81.3%
60-40	75.2%	79.4%	79.4%	67.5%

Additionally, it guarantees that the entire dataset is used for both training and testing without introducing any bias to the results' accuracy. Classification accuracy (Acc) measures how often a classifier makes the right choice. The fraction of wrong decisions is provided by the error rate or Err.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad \ldots\ldots\ldots\ldots\ldots(1)$$

$$Err = \frac{FP + FN}{TP + TN + FP + FN} \ldots\ldots\ldots\ldots\ldots(2)$$

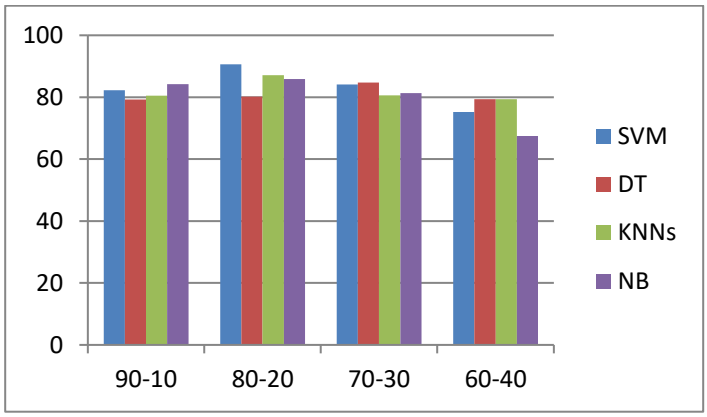

Figure. 2 Show the performance of classifiers

VI. CONCLUSION

In this study, authorship was performed on a collection of English corpus of five different authors A, B, C, D, and E, we divided the features into three groups BoWs, FWs, and stylometric features. We used Naive Bayes (NB) classifier, decision tree, KNN and a support vector machine for the classification tasks. The maximum success rate was set up with an 80 to 20 data ratio, and all mentioned classifiers performed well on our data set. But, the SVM classifier achieved the topmost rate of accuracy, which was 90.6%. We also learned from this trial that combining different stylometry features enhances the result in comparison to using just one type of feature. The resilience of various ML algorithms for jobs involving several authors and small text datasets can be studied in more detail in the future. To expand the scope of our study, we will also incorporate a few fresh pairings of aesthetic elements.

.

ACKNOWLEDGMENT

I am very thankful to my great main supervisor, Dr. Aarti Kumar, cannot be adequately stated in words. This undertaking would not have been possible without the guidance and expertise of Dr. Sanjeev Gupta. Additionally, I am thankful to my colleagues for their assistance, and PANCLEF for supplying the data needed for the study.

REFERENCES

[1] N. Akiva, "Authorship and Plagiarism Detection Using Binary BOW Features", CLEF 2012 Evaluation Labs and Workshop, ISBN 978-88-904810-3-1, 2012. ISSN 2038-4963.

[2] S. Ruseti, and T. Rebedea, "Authorship Identification Using a Reduced Set of Linguistic Features",Notebook for PAN at CLEF 2012

[3] M. Kestemont, K. Luyckx, , W. Daelemans, and T. Crombez, "Cross-genre authorship verification using unmasking. English Studies", 2012, 93(3):340–356.

[4] U.Sapkota, T. Solorio, M. Montes, S.Bethard, and P. Rosso, "Crosstopic authorship attribution: Will out-of-topic data help?", In Proceedings of COLING 2014, *the 25th International Conference on Computational Linguistics: Technical Papers*,2014, pp. 1228–1237.

[5] J.Grieve, "Quantitative authorship attribution :An evaluation of techniques. Literary and Linguistic Computing",2007, 22(3):251–270.

[6] E.Stamatatos, N. Fakotakis, and G. Kokkinakis, "Automatic text categorization in terms of genre and author", *Computational Linguistics*,2000 26(4):471–495.

[7] Savoy, J. "Authorship attribution based on a probabilistic topic model", *Information Processing and Management*, 2013,49(1):341–354.

[8] D. N. Bozkurt,O. Baglıoglu,E. Uyar, "Authorship attribution",2018 https://www.researchgate.net/publication/4321080.

[9] The classic vector space model.http://www.miislita.com/termvector/ term-vector-3.html

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

[11] M.Koppel, J. Schler,S. Argamon, " Authorship attribution in the wild" Language Resources and Evaluation, Advanced Access published,2010, 2010:10.1007/s10579-009-9111-2.

[12] 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.

[13] R. A. J.Matthews, T. V. N. Merriam,," Neural computation in stylometry ", An application to the works of Shakespeare and Marlowe, *Literary and Linguistic Computing*, vol. 8, no. 4,1993, pp. 203–209.

[14] M.Zhang, J. Yao, " A rough sets based approach to feature selection", In: Proc. 23rd International Conference of NAFIPS,2004, pp. 434–439.

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

[16] S. Argamon, C. Whitelaw,P. Chase, " Stylistic text classification using functional lexical features", *Journal of the American Society of Information Science and Technology*,2007, pp. 802–22.

[17] S. Argamon, M. Koppel, Pennebaker, Schler : Automatically Profiling the Author of an Anonymous Text Communications of the ACM , 2013,pp. 119-123.

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

AUTHORS

First Author – Urmila Mahor, MCA,M.Tech (CSE), Ph.D Scholar at Rabindranath Tagore University,
Second Author – Dr. Aarti Kumar, MCA,Ph.D, Rabindranath Tagore University,

.

Correspondence Author – Urmila Mahor,