International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 1, ISSN No. 2455-2143, Pages 197-201
Published Online May 2022 in IJEAST (http://www.ijeast.com)

REVEALING THE TRUE AUTHORSHIP: A CASE STUDY

Urmila Mahor, Aarti Kumar
Department of Computer Science and Engg.
Rabindranath Tagore University, M.P., India

***Abstract*— As digital media is growing, the problem of text proliferation is becoming a big problem. Therefore, the identification of true ownership of a document has become a cumbersome task. In the digital era, it is very easy to copy someone's document and publish it in their name. So it is very necessary to check the true authorship. Authorship attribution becomes difficult when we conduct it manually. However, this process needs automation, when the document size becomes large. AA is a mixture of art, science and technology that helps to discover the genuine authorship of an unknown text/document, based on its specific writing features. These specific features can reflect the author's mood, education, gender, age, ideology, religion, or motivation. Many kinds of characteristics, like lexical, character, structural, syntactic, and semantic are used in authorship recognition. In this experiment, we used approximately 120 different kinds of feature sets. In our experiment, we examined that the logistic classifier was working well and gave good results in the form of accuracy.**

***Keywords*— Authorship attribution, features, machine learning, Weka 3.6.**

I. INTRODUCTION

Authorship attribution determines the authorship of an unknown document by using its literary fashion. Typically, authorship attribution issues can be of two types: closed-set or open-set problems. In a closed set, the nameless document may be of someone inside the group of known authors being examined [18]. In an open set authorship anonymous piece of document may or might not be written by someone inside the set of known authors, or the writer of nameless document can belong to outside the group of authors. Authorship attribution additionally plays a useful character in the forensic technology field. Many researchers proposed a wide variety of style markers to recognize the author's characters of writing [19]. Authorship attribution can play an essential role within the fast and growing electronic text document industry. It is helpful when someone claims his/her authorship for the specific file because stealing electronic documents is common and hassle free. We can say that with this process, a single author's ownership is credited to a portion of the text of an unknown or disputed record.

Authorship Attribution is a composition of art, science and technology, to justify the genuine ownership of a disputed/nameless textual content via comparing the fashion of writing and its features. It recognizes the genuine authorship of an unspecified file via studying the facts and enough evidence of the claimed writer [21].

Authorship attribution falls into the subsequent categories like "Authorship identification" (AI), "Authorship Profiling" (AP), "Authorship Verification" (AV), and "Authorship Clustering" (AC). In the AI, given a cluster of feasible authors or writers for whom a few texts of undisputed authorship exist, the focus of task is to find the genuine writer. The authorship verification approach checks the true ownership of the claimed file, whether it belong to the claimed person or not [23, 24]. Author profiling differentiates authorship with the help of reading the social traits used by writers in their documents. This facilitates author profiling factors like gender, native language, age, emotion, or personality type. In comparison with the classical AA work, authorship clustering is more difficult. Assume we've got a fixed set of different authors' documents, and the work is to create a one-of-a-kind set (or clusters) of documents based totally on each author such that every cluster be associate to a unique writer [22, 1].

II. FEATURES

a. Lexical Features

Another feature is a lexical one. These are some special characters, e.g., letter frequencies, number of content words, spelling mistakes, n-grams as [5] [6], type-token ratio, n-grams [10], sentence length [7], phrase length [8], function words [9, words per phrase type [5], function word ratios, unigrams, word n-grams [3], word bi-grams or sequences, FW frequencies. It can apply in any language and to any corpus with the availability of tokenizers is the main advantage. Another useful type of lexical feature is n-grams, which are denoted as a sequence of consecutive words of length n. Lexical n-grams are becoming popular because of their effectiveness over character n-grams and syntactic n-grams when all the possible identifiers are involved as features [12]. There is a problem with the lexical n-gram approach. With the current set of ngram functions, it is not possible to capture concurrency between words in a longer context due to the limitations of n parameters and the independent assumptions of n-gram.

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 1, ISSN No. 2455-2143, Pages 197-201
Published Online May 2022 in IJEAST (http://www.ijeast.com)

Character and word N-gram

Numerous scientists referenced the lexical highlights utilized in their review with various classifiers and strategies. In a general sense, word-based or character-based highlights are considered lexical features. Lexical elements are word count, length of sentences in words, word length, jargon extravagance, hapax legomena, and hapax dislegomena.

The n-gram addresses are likely the nicest component of the stylometric challenge for small-sized text. Individual scribes have proven to be primarily valued for their language recognition of n-grams [Cavanar and Trenkel, 1994]. Character n-grams have been demonstrated to be feasible for foundation and quantifiable without any extraordinary background facts. The combination of man or woman and stage phrase n-gram capacities offers excellent and specific consequences because of their crucial nature. Character n-grams and phrase n-grams have been chosen as the fundamental highlights in several initiation attribution frameworks [1]. Individual n-grams and phrase n-grams are established as more effective in finding out the initiation of a report by catching the grammar and fashion of its creator. Restrictive style markers like commas, question marks, complete stops, and interjection marks are other fascinating traits that may be confirmed as compelling fashion markers [2]. Regularly, it's been seen that the character bigrams, trigrams, and four-grams with their separate recurrence or tf-idf values are utilized in initiation attribution. Despite the straightforwardness, individual n-grams are extra beneficial and effective [3]. Several researchers used sentence length and male or female n-grams [4] [25] [26] [27] [28] [11]. We believe they are excellent signs.

b. Syntactic Features

Typically regarded as significant and unforgiving semantic highlights, because extremely difficult to intentionally control these features. Often, grammatical form (POS) labels and n-grams highlights are utilized for advent. Grammatical features are more dependable style markers than lexical attributes as they are not under the cognizant control of the writer and can't be controlled intentionally [13] [14]. POS facts are more compelling and might reflect a more solid authorial fingerprint than lexical records. [1]. Syntactic capabilities are solid indicators for authorial attribution. [31]. The benefit of syntactic features is that, they are no longer subject to the writer's conscious manipulation, and thus provide fantastic insights for attribution. [15]. Work phrases are helpful in preventing difficult impersonation or reproduction. A few n-grams are, moreover, handled as word stems. Studies indicate that portrayals of a text archive in mild potential words can be related to records from a low-recurrence degree in dialects, for instance, phrase stems [16] [23]. Those capability words include syntactic perspectives like pronouns (that, they, we, he, she), determiners (that, the), relational phrases (of, in), helper movement words (to be, is), modular (can also, may want to), combination (and, or, however), and quantifiers (some, both). POS records are more common and might reflect reliable authorial fingerprints than lexical information. [15]. Syntactic skills are stable signs of authorial attribution. [17].

c. Structural Features

Primary elements mean how a creator utilizes his or her philosophy to shape a sentence. Now and again, text highlights or structures are imperceptible to the peruse except if we specifically bring them up or learn them. In the underlying highlights, we for the most part center around sentence structure, sentence development, the dynamic or uninvolved voice of the sentence, direct or aberrant discourse utilized in sentence development, the all-out inclusion of sentences in a section, or normal size of the passage in words, sentence length, words per sentence, utilization of extraordinary images and characters, and the way of making sentences are those elements that assist with distinguishing the style.

d. Content-Specific Features

In a particular area or subject, a specific collection of words will take place on a regular basis. These words are called Content-Specific Features. When discussing computers, some words like RAM, ROM, LAPTOP,

DESKTOP will usually be treated with these words as content-specific features. The semantic content of a document is less effective because it is variable in nature, easy to change, and under the conscious control of an authority. While semantic features are difficult to manipulate, they are more useful as compared to content features [13].

III. DATASET CHARACTERSTICS

Traditionally, 10,000 words per author were seen as a reliable minimum for an authoritative set [33]. 10,000 words per author are considered a reliable minimum for an authoritative set [32]. According to a study on text size, it was found that 5,000 words can be considered a minimum requirement in training, and 200 words per author are considered for short text [9]. To fully assess a paternity attribution method, performance must be measured under various conditions [1]. Researchers use more than 10,000 words per author [32] [8] , which is considered a reliable minimum size for an attribution. Some researchers focused on small text sizes, between 500 and 100,200 words per document [32]. If the data size is limited, then attribution becomes difficult because insufficient or insufficient facilities are not able to judge authorship. Traditional approaches are less reliable in this situation [20]. Short texts require robust representations and machine learning (ML) algorithms capable of handling limited data. Reducing the size of training samples directly impacts on classification result and its accuracy. It is very difficult to predict or report a text of a particular length to correctly quantify the stylistic features of an unattributed text.

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 1, ISSN No. 2455-2143, Pages 197-201
Published Online May 2022 in IJEAST (http://www.ijeast.com)

IV. CLASSIFICATION METHODS

4.1 Naive Bayes classifier (NBC)

The Naive Bayes Classifier technique is a Bayesian theorem and is particularly suited when the features vector of input data is high. Despite its simplicity, NBC is often a sophisticated classification method. The NBC builds a probabilistic model for 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 the authors is the author of that anonymous or test text.

Generally, it has been seen that the Naive Bayes classifier is used for attribution of authorship in many languages along with English [28]. The disadvantage of Naive Bayes is that 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. This null count has zero probability, making the NBC unable to predict a class.

4.2 SMO classifier

Sequential Minimal Optimization, or SMO, was proposed by John Platt in 1998. SMO solves a very large quadratic programming (QP) optimization problem with the help of SVM. SMO breaks large quadratic programming problems into a series of smaller possible quadratic programming problems, and then these smallest problems are solved analytically. The SMO required linear memory to train the data set, this makes it easy to handle very large training sets easily.

4.3 Logistic regression classifier (LRC)

Logistic regression (LR) is another method, initially proposed by David Cox in 1958. LR is also another powerful supervised machine learning algorithm. LR is a useful analysis method for classification problems, where you try to determine whether a new sample is the best fit in a category. LR is a useful analytical technique. The main benefit is that it is used for both classification and class probability estimation because it is tied with logistic data distribution. It uses a linear combination of features and applies a nonlinear sigmoidal function on them.

4.4 K-Star classifier

K-Star was developed in 2009 by Husain Aljazzar. K-Star is an instance-based classifier. The K star uses the entropy concept to define its distance metric, it is calculated through the complexity of transforming one instance into another, so it takes into account the probability of this change occurring in a "random walk away" manner.

4.5 Locally Weighted Learning (LWL) classifier

An instance-based algorithm for locally weighted learning (LWL) is used to assign instance weights, which are then used by a specified weighted instance handler. Locally weighted learning classifier algorithms are non-parametric, and the current prediction is predicted by local functions that are used only on a subset of the data. The basic concept of using LWL is that a local model is created based on near neighboring data points of the query point instead of building a global model for the whole function space for each point of interest.

4.6 Decision Trees (J48) 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. In terms of the number of attributes, the number of decision trees that can be constructed is exponential. 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 the 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 the "Gini index" and the "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 drawback of decision trees is that they can be employed in a decision tree when the data contains irrelevant features, resulting in a tree that really is larger than the tree required for classification.

PROPOSED METHOD

Step1: Corpus Collection

Step 2: Pre-processing Step:

2.1 In this step, the corpus is converted to UTF-8 Unicode format.

2.2 In this step numbers, special characters, commas, and full stops are eliminated from the corpus.

2.3 Removed stop words from the corpus but did not use the stemming method on the corpus.

Step 3: Feature Extraction: Extract all the features from the corpus.

Step 4: Feature Selection: Select the useful and most relevant features.

Step 5: Vector Space Model Representation: Calculate Term Frequency (TF) and Inverse Document Frequency (IDF) for every document from Step 4 and represent all the documents of the authors as Vector Space Model.

Step 6: Document Generation: Generate document vectors with document weights to build a classification model.

Step 7: Classification: The classification model is used to predict the author of an unidentified document.

Step 8 : Comparison and Result: Compare the outcome of different classifiers and prepare a result based on accuracy.

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 1, ISSN No. 2455-2143, Pages 197-201
Published Online May 2022 in IJEAST (http://www.ijeast.com)

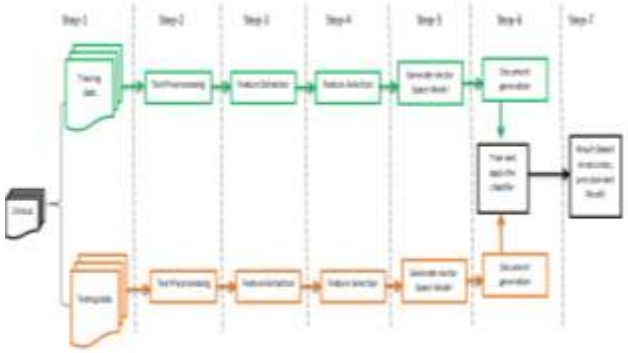

Fig 1. Process of Authorship Attribution

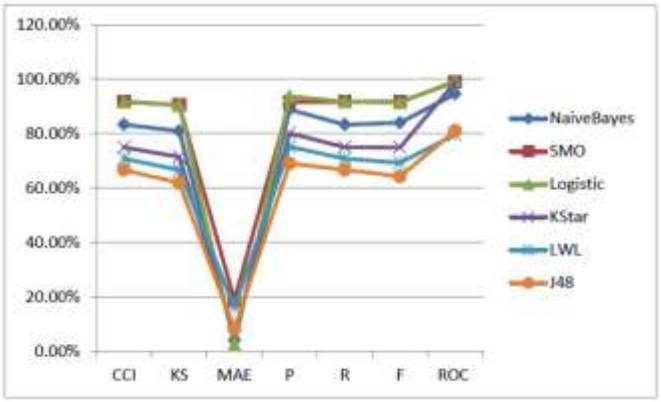

Fig 2. Graphical representation of classifiers

V. EXPERIMENT AND RESULT

For our experiment, we used the Weka 3.6 tools. Cross-validation is a classification procedure, and we employed 10-fold cross-validation, in which data is evenly divided into 10 data sets. The data is trained with nine partitions and tested with one remaining data set. The whole process is repeated 10 times to train the classifier model. The average of all the accuracy estimates obtained across these iterations is then combined into a single accuracy estimate. By using this validation approach, more data is used to train the models.
Additionally, it ensures that all the data in the dataset is used for both training and testing without causing any bias toward the accuracy of the results.

PAN-CLEF Database

For this experiment, we collected the data from PANCLEF, which is an official website that provides data for research purposes. In this data, we have a collection of documents written by different authors. The training data represents novel-length samples of works by the authors named in the file names. This data also contains testing data.

VI. RESULT AND CONCLUSION

Table 1 shows the classification result

Classifier name	Correctly Classified Instances	Kappa statistic	Mean absolute error	Precision	Recall	F-Measure	ROC Area
Naive Bayes	83.30%	81%	4%	89%	83%	84%	95%
SMO	91.67%	90%	19%	92%	92%	92%	99%
Logistic	**91.67%**	**90%**	**2%**	**94%**	**92%**	**91%**	**99%**
KStar	75%	71%	6%	80%	75%	75%	99%
LWL	70.83%	67%	16%	75%	71%	69%	80%
J48	66.67%	62%	8%	69%	67%	64%	81%

CCI-Correctly Classified Instances, KS- Kappa statistic, MAE- Mean absolute error, P- Precision, R- Recall, F-Measure, ROC- ROC Area

In our experiment, we used the PANCLEF data set for authorship attribution. We used the Weka toolkit to perform an experiment on our dataset. We applied six classifiers to classify our data set and used a 10-fold cross-validation approach. The results of various classifiers were compared, and we see that the logistic classifier and SMO classifier perform well on our data set. The logistic classifier is performing better in terms of accuracy and lowering MAE in our experiment.

VII. REFERENCE

[1]. E.Stamatatos, "A survey of modern authorship attribution methods". Journal of the American Society for Information Science and Technology, 2009.

[2]. A. Johnson and David Wright, "Identifying idiolect in forensic authorship attribution: an n-gram text bite approach", anguage and Law / Linguagem e Direito, vol. 1, pp. 37-69,2014,

[3]. J. Grieve, "Quantitative authorship attribution: an evaluation of techniques", Literary and Linguistic Computing, 22(3): pp.251–70,2007.

[4]. E. Stamatatos, "Author identification using imbalanced and limited training texts", Proceedings of the 18th International Conference on Database and Expert Systems Applications, Regensburg, Germany: IEEE Computer Society, pp. 237–241, 2007.

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

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

[7]. S. Argamon, M. Saric, and 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.

[8]. M. Gamon, "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.

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

[10]. F. Peng, , D. Schuurmans, V.Keselj, and S. Wang, "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. 2003.

[11]. D. V. Khmelev, F. J. Tweedie, "Using Markov chains for identification of writers. Literary and Linguistic Computing, 16(4), pp. 299-307,2002.

[12]. Ding, B. C. M. Fung, and D. Mourad , " A visualizable evidence-driven approach", ACM Transactions on Information and System Security, vol. 17, No. 3, Article 12, 2015.

[13]. H. Baayen, V. H. Halteren, and F. Tweedie, "Outside the Cave of Shadows: Using syntactic annotation to enhance authorship attribution", Literary and Linguistic Computing, 11(3), pp. 121–31, 1996.

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

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

[16]. J. Burrows, "All the way through testing for authorship in different frequency strata", Literary and Linguistic Computing, 22(1),pp. 27–47, 2007.

[17]. Zheng, L.,Jiexun, Chen, and Huang, " A framework for authorship identification of online messages: Writing-style features and classification techniques", J. Am. Soc. Inf. Sci. Technol., 57(3), pp. 378–393, 2006.

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

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

[20]. Zheng, H. Chen and Z. Huang ,"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), pp.378-393, 2006.

[21]. E.Stamatatos, N. Fakotakis, and Kokkinakis , "Computer-based authorship attribution without lexical measures" , Computers andthe Humanities, 35(2), pp. 193-214, 2001.

[22]. E. Stamatatos, M. Tschuggnall, B. Verhoeven, W. Daelemans, S. Gunther, V. Ben and M. Potthast,"Clustering by authorship within and across documents" Working Notes Papers of the CLEF,2016.

[23]. M. Koppel, J. Schler, and , S, Argamon, " Authorship Attribution: What's Easy and What's Hard?",2013.

[24]. E. Stamatatos, B. Verhoeven, W. Daelemans, S. Gunther, V. Ben and M. Potthast, "Overview of the Author Identification Task at PAN 2014." CLEF , 2014.
[25] B. Kjell, " Authorship determination using letter pair frequencies with neural network classifiers", Literary and Linguistic Computing, 9(2), pp. 119-124.

[25]. B. Kjell, W. A. Woods, O. Frieder, "Information retrieval using letter tuples with neural network and nearest neighbor classifiers", In IEEE International Conference on Systems, Man and Cybernetics, vol. 2, pp.1222-1225, 1995.

[26]. G. R. Ledger, and T. V. N. Merriam, "Shakespeare, Fletcher, and the Two Noble Kinsmen", Literary & Linguistic Computing 9,pp. 235-248, 1994.

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