Authorship Attribution using Content based Features and N-gram features

Raju Dara, T. Raghunadha Reddy

Abstract*: The internet is increasing exponentially with textual content primarily through social websites. The problems were also increasing with anonymous textual data in the internet. The researchers are searching for alternative techniques to know the author of an unknown document. Authorship Attribution is one such technique to predict the details of an unknown document. The researchers extracted various classes of stylistic features like character, lexical, syntactic, structural, content and semantic features to distinguish the authors writing style. In this work, the experiment performed with most frequent content specific features, n-grams of character, word and POS tags. A standard dataset is used for experimentation and identified that the combination of content based and n-gram features achieved best accuracy for prediction of author. Two standard classification algorithms were used for author prediction. The Random forest classifier attained best accuracy for prediction of author when compared with Naïve Bayes Multinomial classifier. The achieved results were good compared to many existing solutions to the Authorship Attribution.

Keywords* : Authorship Attribution, Accuracy, N-grams, Author Prediction, Content based features.

I. INTRODUCTION

The internet is increasing exponentially with the enormous quantity of text day by day through blogs, reviews, tweets and other social data content. The information analysis needs to undergo certain processes for execution of text, understanding the text and streaming of the text. The researchers need the automated tools to process the information which is dynamic in nature. In this process, sometimes it is necessary to identify the owner who has created the text or document. Authorship Analysis is one area used by the researchers to find the author details of the text.

In general the Authorship Analysis is made as three types namely Authorship Verification, Authorship Profiling and Authorship Attribution [1]. The Authorship Verification is a process of comparing multiple chunks of written text of a particular author to identify whether the text was written by the same author or not [2]. Authorship Profiling is used to predict the profiling characteristics such as age, location, gender, personality traits, native language, occupation, educational background by examining the authors style of writing [3]. In authorship attribution, given a Document (anonymous), it is to be compared with the documents of different authors. That means 'n' no. of authors along with their writings is compared with anonymous text and there by identifies the real author [4].

In authorship attribution, two problems are addressed mainly, closed set and open set problem. In the closed set problems all documents are considered as a set, called candidate author set and when author of unknown text is belonging to the candidate authors set then it is called closed set problem, otherwise the unknown text not belongs to any one of the author included in candidate authors set then it is called open set problem. The Authorship Attribution is used in various applications like literary research, security aspects and forensic analysis [5]. The property wills and suicide notes are examined using this authorship attribution where a will or note is written by suspected author or not, by considering the suspected authors style of writing. The threatening mails from the terrorist organizations undergo at authorship attribution techniques to verify whether they originated from the correct source or not. Steeling of information by wrongly claiming about certain innovations is more often observed in literary research where in the authorship attribution is used to analyze the authors style of writing.

An Authorship Attribution framework significantly developed from the three segments like selection of feature, representation of text and training. Extraction of the key textual features has been focused by most of the researchers to build the efficient classification model [6]. The researcher was identified different combinations of features to distinguish the authors style of writing. Some of them are character, lexical, structural based features, content based features and readability features. The number of features applied on the classification plays a vital role in the document representation. Efficient classification models were constructed with effective feature selection algorithms which involves dimensionality reduction. Text documents are represented in the form vectors that were easily understand by the classification algorithms. A supervised learning algorithm is used to construct a classifier, it is used for training this classifier and it is utilized in assigning the class labels for new documents. In this work, Bag of words approach is used for representing the documents. In this model, every document is represented with features like tokens, terms, phrases and the related weights are calculated based on the weight measures.

This paper is organized in 6 sections. Section 2 reviewed the existing work of Authorship Attribution. The dataset characteristics were presented in section 3. The traditional model for representing the document vectors and the features used in this work were described in section 4. The empirical results of this work were analyzed in section 5. The conclusions and future work

Revised Manuscript Received on October 15, 2019

Dr. Raju Dara, Computer Science and Engineering, Vignana Bharathi Institute of Technology, Hyderabad, Telangana, India.
Email: rajurdara@gmail.com

***Dr. T. Raghunadha Reddy**, Information Technology, Vardhaman College of Engineering, Hyderabad, Telangana, India.
Email: Raghu.sas@gmail.com

were specified in section 6.

II. RELATED WORK

Information extraction from different written text documents to recognize the facts about the authors had become a subject of interest in the recent time. Several researchers proposed many successful approaches to Authorship Attribution. These approaches were categorized based on the stylistic features used to distinguish the writing styles of different authors and the data set which is involved to do it. Generally any information retrieval problems like Authorship Attribution involves the following steps in finding the solution such as preprocessing, feature extraction and applying the learning methods [3].

Preprocessing plays vital role in information extraction and text processing. In view of representing the data in the machine format, preprocessing is necessary because the raw data may be available in different formats, sometimes few values may be missing, to maintain the similarity preprocessing applied to the data set. It is the same case with the Authorship attribution approaches. In order to create the good environment for data analysis and to attain the reduced complexity, data preprocessing is mandatory. Various techniques and tools were used by many researchers to preprocess the data for Authorship Attribution. Most of the researchers used different type of preprocessing techniques like elimination of HTML code in the tweets [7], removal of mentions, hash tags and URL's [8], changed the entire text in to lower case [8], elimination of the emojis [9], removed multiple white spaces and invalid characters [9].

Stylometry works on the assumption that every author has specific style of writing and it has some specific features. These features provide a ground to identify the author. In general the features of stylometry are classified into character based, lexical, syntactic and semantic features. Generally text is viewed as string of characters. Some of the character-based features are number of letters, uppercase characters, digits, white spaces, special characters, and no. of occurrences of each letter. This type of features consisting of the rules used to form sentences like function words, punctuation. Usage pattern of function word is a useful feature for authorship identification. In this way, different character level measures were characterized, it includes digit count, alphabet count, count of lowercase and uppercase characters, frequency of letter, count of punctuation marks [10]. The character level n-gram features are important in dealing with the character based features. The most repeatedly occurring character n-grams will play major role in stylistic purposes. Numerous varieties of tools are not required to attain most repeated n-grams, and attainment process is fully independent of language used. However, Stamatatos et.al., addressed [3] that when compared to word-based approach, degree of representation is substantially raised. The reason is very clear that n-grams will catch up unessential information and no. of character n-grams are require to symbolize a unique lengthy word. Magdalena et.al., considered [11] frequency of the most common 4-grams character. In the work of Erwan Moreau [12] considered Character uni-grams, tri-grams and penta-grams for characterizing the text.

In general word based features are considered as Lexical features. Some of the Word-Based features are count of all words, count of words in a Sentence, length of the word and Vocabulary Richness, these metrics contains number of words which appears only one time called as hapax legomena and appears two times is called as hapax dislegomena. Different type of lexical features are special characters, letter frequency, content words, misspellings, Verbal Phrases [3], phrase length [9], function words [11], words per phrase type, phrase types [12], function word-token ratios, type-token ratio, unigrams, word n-grams [13], words bigrams or sequences, Function word frequencies, POS trigrams, Pos-Bigrams, Pos-Trigrams [14], Complexity measures with Pos [14], Function words [1], non-function words [3].

In English language, the function words are having considerably less meaningful content and these terms thought of as structured grammatical terms and include a structural relationship with different terms during formation of a sentence. This type of function words comprises of grammatical aspects of English like conjunctions, determiners, prepositions, modals, pronouns, auxiliary verbs and quantifiers. Gilad Gressel et.al., retrieved [15] seven features from the text document and these features comprises of grammatical characteristics like pronouns, adjectives, nouns, adverbs determiners and foreign words. Based on the contextual information, for every word morpho syntactic information tags were allocated and which is a technique handled by Part of speech (Pos) Tagger [16].

Structural features appear in managing the organization of text and its outline called structure. Researchers usually concentrates on structure of the words such as good wishes signatures, the total count of paragraphs and average length of the paragraph, spam detection conversation length [9], Average sentence length, the HTML tags [14], the URL's, the set of common slang vocabulary and the emoticons [15].

In a Particular domain topic, a specific set of words will come on a regular basis those words are called Content-specific features. While discussing about computers some words like RAM, ROM, LAPTOP and DESKTOP will appear, these words are treated as content specific features.

III. DATASET CHARACTERISTICS

In this work, the experiment performed on PAN competition 2014 dataset for Authorship Attribution. The dataset characteristics were presented in table I. the researchers used various measures to evaluate the efficiency of the Authorship Attribution system. In this work, accuracy measure is used to test the efficiency of our approach. Accuracy is the number of test documents were correctly predicted their author.

Table- I: The Characteristics of dataset

Features	Training data	Testing data
Total documents	500	100
Total authors	100	100
Total Documents per author	5	1
Total words	41583	12764
Average number of words per document	1135	1121
Average number of words per sentence	25	21

IV. BAG OF WORDS (BOW) MODEL

The BOW model is considered in the recent studies as a widely used document representation model wherein there are different other models such as continuous space model, word2vec model, doc2vec and network based model. The frequency of a word in a document is considered to be a feature in the BOW model whereas the other models do embed the word/document in a named vector space where the dimensions of the vector space is considered as the semantic similarity of features. In network model, the document is represented as a graph with words as vertices and the relationship among words in sentences is also considered in the representation.

There are two categories of document representation models wherein the first category contains the features which are designed at words level and in the second representation model the features are at the total document level. The model BOW represents first category and the words frequency is represented as feature values. Similarly the word2vec model is also belongs to first category wherein it embeds the word in vector space. In the second category which is similar to that of the doc2vec model relies on the document level which embeds the total document as a vector. Finally the network based model belongs to both categories which quantify the properties of the network nodes belongs the first category wherein the other quantifies the entire network properties corresponds to the second category. In the subsequent sections detailed explanation is made which describe the BOW model and its applicability on authorship attribution approach.

The BOW model is the most standard approach which is adapted by many researchers in various text processing, information retrieval and text classification domains. In this model every document is represented as unordered features set which corresponds to the vocabulary is in terms of words, word sequence (token n-grams), POS n-grams and sequence of letters up to a length n (Character n-grams). The term of a vocabulary is a numerical value. Most commonly used feature value identification is term frequency (TF) and TF-IDF. While calculating the frequency, in TF, no. of occurrences of a specific word is to be calculated and where as in TF-IDF the reciprocal document frequency set is multiplied with the term frequency (TF). By taking this multiplication we can improve the importance of rare terms and can reduce the importance of the term which occurs in many documents [17]. Hierarchically in all engineering applications, the BOW model has emerged because of its simplicity and sometimes it results high accuracy among the document representation models.

In this work, two types of features such as content specific features and n-grams were considered for representing the vectors of documents. Fig. 1 Shows the BOW model for representing the document vectors with content specific features.

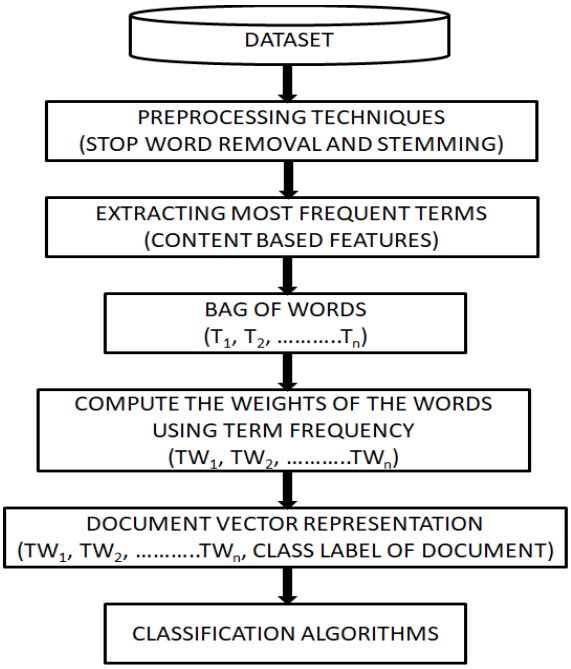

Fig. 1. The Bag Of Words model for extracting content based features

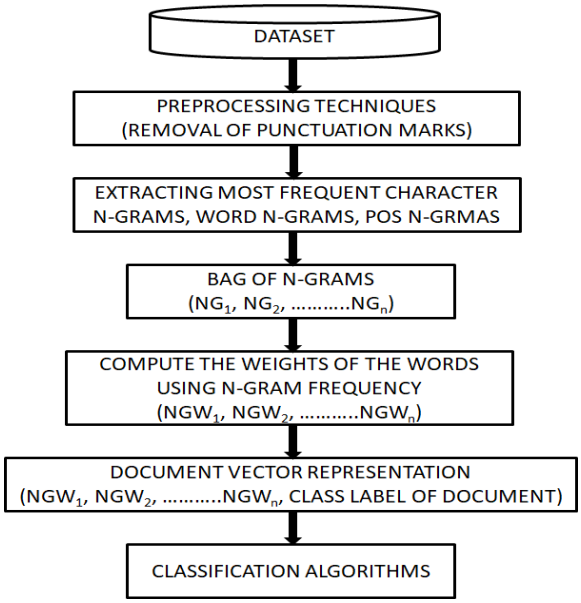

Fig. 2.The Bag Of Words model for extracting n-gram features

In fig. 1. First, we performed preprocessing techniques such as stopword elimination and stemming on the dataset. Then extract the most frequent content specific terms and considered as bag of words. These bag of words were used to represent the document vector. Classification algorithms take these vectors and generate the classification model.

The fig. 2 Shows the BOW model for representing the document vectors with most frequent n-grams of character, word and POS tags where n rage is from 1 to 3. Consider the n-grams as bag of words. In this model initially applied with the preprocessing technique such as removal of punctuation marks and extracts the n-gram features which reflects the authors style of writing and is having the differentiating power to compare the authors style of writing. Every document in the BOW is represented as the bag of word and the weight of each BOW is represented as a value. In the author attribution the frequency of the BOW is considered to represent the document vector.

Every document vector contains the numerical weights of features or terms extracted from the dataset. Authorship Attribution mainly depends on the weights of the features in the document but the process of calculating the term weight certainly affects the accuracy of classification. In this work, we used frequency of content specific features and frequency of n-gram for representing the document vector.

V. EMPIRICAL RESULTS

Classification is a process of assigning a predefined class label to an unknown document. To classify the documents, classification algorithms were used. In this work, two classification algorithms such as Naive Bayes Multinomial (NBM) and Random Forest (RF) were used to test the efficiency of author prediction. In this algorithms, 10-fold cross validation is used where in the dataset is randomly divided into 10 samples. In each iteration, 9 samples were used for training the classifier and 1 sample is used to test the efficiency of the trained classifier. This process is repeated till every sample is used for testing the trained classifier.

In this work, first we extracted 4000 most frequent content based features as a bag of words. Each document is represented with these 4000 words as document vectors. The classification algorithms used these vectors to generate classification model. The accuracies of the classifiers for author prediction are presented in table II.

Table- II: The accuracies of author prediction with most frequent content based features

Classifier/ Number of features	NBM	RF
1000	64.22	67.56
2000	66.07	69.78
3000	67.81	71.35
4000	69.39	75.91

In table II, the RF classifier attained 75.91% accuracy for prediction of author when most frequent 4000 content based words are used as features. The performance of RF classifier is good when compared with NBM classifier. It was detected that the accuracies are improved with increase in the count of words.

Later, the experiment continued with most frequent character unigrams, character bigrams, character trigrams, word unigrams, word bigrams, word trigrams, POS unigrams, POS bigrams and POS trigrams. We identified 2000 most frequent n-grams as bag of words. Each document is represented as document vector with these 2000 n-grams. The accuracies of these n-grams for author prediction are represented in table III.

Table- III:. The accuracies of author prediction with most frequent character, word and POS n-grams

Classifier/ Number of features	NBM	RF
500	70.81	74.47
1000	73.74	76.19
1500	75.54	81.84
2000	80.39	84.12

In table III, the RF classifier obtained 84.12% accuracy for prediction of author when most frequent most frequent character, word and POS n-grams were considered as features. The performance of RF classifier is good when compared with NBM classifier in all iterations. It was observed that the accuracies are increased with the increase in count of n-grams. It was also witnessed that the performance of n-grams is good for prediction of author when compared with the accuracies content based features.

Finally, we experimented with the combination of content based features and character, word and POS n-grams. In this experiment, the content based features are fixed in every iteration and the n-grams are changed from 500 to 2000. The accuracies of the combination of features for author prediction is represented in table IV.

Table- IV: The accuracies of author prediction when combination of content based features and n-grams were used

Classifier/ Number of features	NBM	RF
4000 terms + 500 POS	82.18	85.23
4000 terms + 1000 POS	83.47	87.71
4000 terms + 1500 POS	85.91	88.67
4000 terms + 2000 POS	86.78	91.87

The Random Forest classifier got accuracy of 91.87% which is depicted in the table IV in classifying author when the document is represented with most frequent 4000 content based features and 2000 n-grams and also observed that n-gram features and content based features alone are not improving the prediction accuracy of an author. The performance of RF classifier is best when compared with the performance of NBM classifier in all iterations. It was also found that the accuracies were increased with the increase in count of features. We continued the experiment with more than 4000 content based features and more than 2000 n-gram features and it was observed that the reduction in accuracies of author prediction.

VI. CONCLUSION AND FUTURE WORK

In this work, the RF classifier obtained good results for author prediction in PAN 2014 competition dataset. The experiment performed with content based features of 4000 and most frequent character, word, POS n-grams features of 2000. It was observed that the performance of the n-grams was good for prediction of author when compared with content based features. It was also found that there is an improvement in accuracy when experimented with the combination of content based and n-gram features. The Random forest classifier obtained 91.87% accuracy when experimented with the combination of features.

In future work, we concentrated on finding new weight measures for representing the document vector thereby improving the accuracy of the author prediction. It was planned to use deep learning techniques in Authorship Attribution to improve the author prediction accuracy.

ACKNOWLEDGEMENT

I will be grateful to the FIST laboratories for providing required computational facilities at the institution, and I extend my deep sense of gratitude to the management, principal, director of R&D of Vignana Bharathi Institute of Technology for the accorded support.

REFERENCES

1. J. Schler, Moshe Koppel, S. Argamon and J. Pennebaker (2006), Effects of Age and Gender on Blogging, in Proc. of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, March 2006. Vol. 6, (2006), 199-205.
2. Raghunadha Reddy T, Vishnu Vardhan B, Vijayapal Reddy P, "A Survey on Author Profiling Techniques", International Journal of Applied Engineering Research, March 2016, Volume-11, Issue-5, pp. 3092-3102.
3. 3. Stamatatos, E. (2009). A Survey of Modern Authorship Attribution Methods. Journal of the American Society for Information Science and Technology, 60(3), 238-556.
4. Raghunadha Reddy T, Vishnu Vardhan B, Vijayapal Reddy P, "Profile specific Document Weighted approach using a New Term Weighting Measure for Author Profiling ", International Journal of Intelligent Engineering and Systems, 9 (4), pp. 136-146, Nov 2016.
5. M. Sreenivas, Raghunadha Reddy T, Vishnu Vardhan B, "A Novel Document Representation Approach for Authorship Attribution", International Journal of Intelligent Engineering and Systems, 11 (3), pp. 261-270, MAY 2018.
6. Raghunadha Reddy T, Vishnu Vardhan B, Vijayapal Reddy P, "Author profile prediction using pivoted unique term normalization", Indian Journal of Science and Technology, Vol 9, Issue 46, Dec 2016.
7. Ludovic Tanguy, Franck Sajous, Basilio Calderone, and Nabil Hathout. Authorship attribution: using rich linguistic features when training data is scarce, CLEF 2012 Evaluation Labs and Workshop, 17-20 September, Rome, Italy, September 2012. ISBN 978-88-904810-3-1.
8. Koppel, M., Schler, J., Bonchek-Dokow, E.: Measuring differentiability: Unmasking pseudonymous authors. J. Mach. Learn. Res. 8, 1261–1276 (Dec 2007).
9. Navot Akiva. Authorship and Plagiarism Detection Using Binary BOW Features, CLEF 2012 Evaluation Labs and Workshop, 17-20 September, Rome, Italy, September 2012. ISBN 978-88-904810-3-1. ISSN 2038-4963.
10. Lucie Flekova and Iryna Gurevych. Can We Hide in the Web? Large Scale Simultaneous Age and Gender Author Profiling in Social Media—Notebook for PAN at CLEF 2013.
11. Magdalena Jankowska, Vlado Keselj, and Evangelos Milios. CNG Text Classification for Authorship Profiling Task—Notebook for PAN at CLEF 2013.
12. Erwan Moreau and Carl Vogel. Style-based Distance Features for Author Profiling—Notebook for PAN at CLEF 2013.
13. Raghunadha Reddy T, Vishnu Vardhan B, Vijayapal Reddy P, "N-gram approach for Gender Prediction",Â 7 th IEEE International Advanced Computing Conference, Hyderabad, Telangana, PP. 860-865, Jan 5-7, 2017.
14. Juola, P.: Authorship attribution. Found. Trends Inf. Retr. 1 (2006) 233-334
15. Gilad Gressel, Hrudya P, Surendran K, Thara S, Aravind A, and Prabaharan Poomachandran. Ensemble Learning Approach for Author Profiling—Notebook for PAN at CLEF 2014.
16. Stefan Ruseti and Traian Rebedea. Authorship Identification Using a Reduced Set of Linguistic Features—Notebook for PAN at CLEF 2012. CLEF 2012 Evaluation Labs and Workshop, 17-20 September, Rome, Italy, September 2012. ISBN 978-88-904810-3-1. ISSN 2038-4963.
17. Raghunadha Reddy T, Vishnu Vardhan B, Vijayapal Reddy P, "A Document Weighted Approach for Gender and Age Prediction", International Journal of Engineering -Transactions B: Applications, Volume 30, Number 5, pp. 647-653, May 2017.

AUTHORS PROFILE

Dr. Raju Dara is a professor of Computer Science and Engineering Department at Vignana Bharathi Institute of Technology, Hyderabad. He has 16 years of teaching experience for Graduate and Post Graduate engineering courses. His current research interests are Data Warehousing, Image Processing, and Network Security. He published 30 research papers in international journals as well as international conferences.

Dr. T. Raghunadha Reddy, working as Associate Professor in the Department of Information Technology, Vardhaman College of Engineering, Shamshabad, Hyderabad, Telangana, India. He has 15 years of teaching experience. His current research interests are Data mining, Natural Language Processing, Information Retrieval and Text Classification. He published 40 research papers in reputed international journals and international conferences. He has memberships in IET and IAENG.