Received: February 12, 2018 261

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

A Novel Document Representation Approach for Authorship Attribution

Sreenivas Mekala[1]* Raghunadha Reddy Tippireddy[2] Vishnu Vardhan Bulusu[3]

[1]*Department of Information Technology, Sreenidhi Institute of Science and Technology, Hyderabad, India*
[2]*Department of Information Technology, Vardhaman College of Engineering, Hyderabad, India*
[3]*Department of Computer Science and Engineering, JNTUH, Jagtiyal, Karimnagar, India*
* Corresponding author's Email: msreenivas@sreenidhi.edu.in

Abstract: The rapidly growing data in the web result in stolen, unidentified and fraudulent data. Identification of such data is of a prime objective for forensic departments, researchers and governments. In this context, authorship analysis is very useful to reveal the truth by analyzing the text. Authorship analysis is observing the properties of a text to predict authorship of a document. Stylometry is the root for authorship analysis, which is a linguistic research field that exploits the machine learning techniques as well as knowledge of statistics. Authorship Attribution is a type of authorship analysis technique, which is aimed at recognizing the author of an anonymous text within a closed set of authors or subjects. Most of the researchers in Authorship Attribution approaches proposed various set of stylistic features to differentiate the authors based on style of writing. It was observed from the literature the accuracy of author prediction was not satisfactory with stylistic features. In this paper, the experimentation carried out with various stylistic features, feature selection measures and term weight measures identified in various text processing domains to predict the author of a new document. A new document representation approach is proposed to improve the prediction accuracy of author prediction. In the proposed approach the documents were represented with the weights of the documents specific to author group of documents. The results show that the proposed approach obtained good accuracies when compared with the results of stylistic features, feature section measures, term weight measures and most of the existing approaches.

Keywords: Authorship attribution, Stylistic features, Feature selection algorithms, Term weight measures, Document weight measure, Classification algorithms.

1. Introduction

All manuscripts must be in English. These guidelines include complete descriptions of the fonts, spacing, and related information for producing your proceedings manuscripts.

The World Wide Web is increasing rapidly with textual information along with the cyber crimes are also increased in the WWW. The people are sending harassing messages in social media and the terrorist organizations send threatening mails without specifying their correct authorization details. In this context, the researchers are attracted to know the details of the authors of these texts. Authorship Analysis is one such area to predict the author information of a text by analyzing the writing styles of the authors. Various applications such as security, research literature and forensic analysis were used Authorship Analysis techniques.

Authorship Analysis technique was categorized into three types such as Authorship Verification, Authorship Attribution, Authorship Profiling [1]. Authorship Attribution is used to predict the author of a document by processing the texts of several authors [2]. Authorship verification is used to verify whether the document was written by the suspected author or not by analyzing the texts of suspected author [3]. Authorship profiling is used to predict the demographic characteristics of the authors [4]. In this work, Authorship attribution technique is concentrated to predict the author of anonymous documents. Most of the researchers proposed various types of stylistic features to differentiate the

authors writing styles in Authorship Attribution. The general approach followed in most of the existing approaches of Authorship Attribution was BOW (Bag Of Words) approach for document representation.

In this work, the experimentation starts with different types of stylistic features. It was identified that the stylistic features were not sufficient to improve the accuracy of author prediction. Later most frequent terms were extracted from the corpus. The results of BOW model with most frequent terms were not satisfactory. Then, feature selection measures were used to find the informative terms from most frequent terms. The experimentation continued with the features identified by the feature selection algorithms. Different term weight measures from various domains were evaluated to test the impact of term weight measures in author identification. It was observed that the results of term weight measures are good for author prediction.

Finally, a novel approach namely weighted document approach for authorship attribution was proposed to increase the author prediction accuracy. The proposed approach achieved best results for author identification when compared with the results of stylistic features, results of features selection algorithms and the results of term weight measures. In the proposed approach, a new term weight measure is proposed to compute the weights of the terms. A new document weight measure is used to compute the document weight and the document vectors were generated with these document weights. The stylistic features and most frequent terms are independently participated in the classification process but they are collaboratively participated in the proposed approach. This is the reason the proposed approach obtained best results for author identification.

This paper structured in 8 sections. The related work in Authorship Attribution is explained in section 2. The reviews dataset characteristics and evaluation measures used to evaluate the classifiers were explained in section 3. Section 4 explains the basic document representation technique BOW model and also present the experimental results of stylistic features and most frequent terms using BOW model. The importance of feature selection measures and the experimental results of features identified by the feature selection measures were discussed in section 5. The analysis of various term weight measures and experimental results were explained in section 6. Section 7 describes a novel document representation technique namely weighted document specific to author for author identification and also presented the experimental results of proposed approach. The conclusions and future scope is explained in section 8.

2. Literature survey

The style of writing is a primary indicator of an individual identity to predicting the author of a document in Authorship Attribution. In general three steps followed in Authorship Attribution approaches. First, the most discriminative features were identified to differentiate the authors writing styles. Second, the document representation models were identified to represent the document with these features. Finally, the suitable machine learning classification algorithms were detected to predict the author of an anonymous document [5].

Most of the researchers used stylistic features to differentiate the writing style of the authors in Authorship Attribution. Ludovic Tanguy et al., extracted [6] rich set of language specific features like contracted forms, character trigrams, POS trigrams, lexical generosity and ambiguity, phrasal verbs, syntactic complexity, syntactic dependencies, lexical cohesion, lexical absolute frequency, morphological complexity, quotations, punctuation, first/third person proper and narrative. They noted that the performance of set of rich linguistic features was better for author prediction when compared with word frequencies and trigrams of characters. Another researchers obtained [7] best results when combination of word based and character tetra-grams features are used. In [8], the researchers extracted POS bigrams and trigrams, character trigrams, percentage of direct speech from the documents and syntactic features. They obtained overall accuracy of 77% in Authorship identification and found that the author prediction accuracy was improved when the application specific features were added to existing feature set.

The classification algorithms also play an important role in the performance of author prediction. Darnes Vilarino experimented [9] with three supervised learning methods such as Naïve Bayes, rocchio and greedy. It was observed that the rocchio method perform well compared to naïve bayes and greedy methods. George k. mikros et al., extracted [10] character bigrams, character trigrams, word unigrams, word bigrams and word trigrams features from e-mail corpus. They obtained best results when logistic regression and one class machine learning methods were used for author prediction.

Some researchers used different types approaches for analyzing the writing style changes of the authors. Rexha et al., adopted [11] a text

segmentation algorithm to predict the author changes in the dataset of PubMed articles. They used set of stylistic features to identify the authors style of writing and observed that their approach identified more number of writing style changes when the article was written by more number of authors. Another researcher used [12] probabilistic context free grammar for predicting author of a new document. In this approach, probabilistic context free grammar constructed for every author and used this grammar for classification.

The researchers used different types of document representation techniques for author prediction. N. Akiva used [13] binary Bag of Words representation to represent the document vector, which captures absence or presence of common words in a document. It was identified that the author prediction accuracy was improved when the number of texts was increased in the training data. Whereas another researcher proposed [14] a document occurrence representation for author prediction and observed that their representation outperforms when compared with Bag of Words approach and also observed that this document representation works good for small data sets.

3. Dataset characteristics and evaluation measures

3.1 Dataset characteristics

The dataset was collected from amazon.com and it contains 10 different authors reviews on different products. The corpus is balanced in terms of number of documents in each author group and each author group contains 400 reviews of each.

3.2 Evaluation measures

Various measures are used such as precision, recall, F1 measure and accuracy by the researchers in Authorship Attribution to test the accuracy of author prediction. In this work, accuracy measure is used to evaluate the performance of the author prediction. Accuracy measure is the ratio of number of documents correctly predicted their author to total number of documents

4. BOW model

The design of BOW model is depicted in Fig. 1. In this model, preprocessing techniques such as stopword removal and stemming were applied on the dataset to remove the terms which are weak in text discrimination. The features were extracted from the updated dataset. Treat these most frequent terms as bag of words. The documents were represented with this bag of words. The term frequency was considered to represent the weight of the terms in document vectors. In this work, different term weight measures were identified to assign weight to the terms. The machine learning classifiers were used to produce the classification model.

In this work, the experimentation starts with various types of 39 stylistic features. Table 2 shows the different types of stylistic features used in our work.

The document vectors were represented with these 39 stylistic features. Different classification algorithms such as Simple Logistic (SL), Logistic (LOG), IBK, Bagging (BAG), Random Forest (RF) and Naïve Bayes Multinomial (NBM) classifiers were used to create the classification model. The accuracies of author prediction using stylistic features are presented in table 3.

The Random Forest classifier obtained good accuracy of 60.23% compared with other classifiers for author prediction. It was observed that the stylistic features are not more suitable to predict the author of a document.

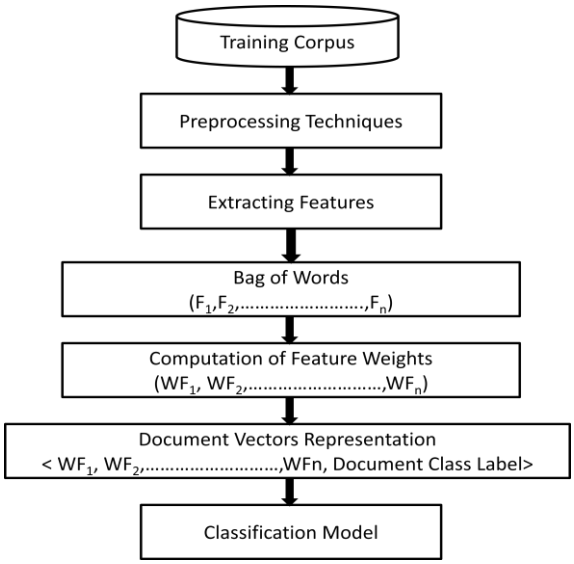

Figure.1 The procedure of BOW model

Table 1. The accuracies of author prediction when stylistic features used

Features/ Classification Algorithms	39 Features
NBM	58.41
SLOG	54.38
LOG	56.87
IBK	55.92
BAG	52.76
RF	60.23

Table 2. Stylistic Features

S no	Stylistic features	Example Features
1	character based features	Characters count
		Count of small letters
		Count of punctuation marks
		ratio of capital letters to small letters
		Ratio of white spaces to non white spaces
		Ratio of numeric characters to non numeric characters
		Ratio of white space to total number of characters
		Ratio of tab spaces to total number of characters
		Ratio of capital letters to total number of characters
2	word based features	Count of words
		Capital letters words
		Count of positive words
		Count r of negative words
		Average word length
		Contraction words
		The ratio of number of words length greater than six to total number of words
		The ratio of number of words length less than three to total number of words
		Count of words with hyphens
		Number of words length greater than 6
		Words followed by digits
		Unique terms
		Ratio of number of words which contain more than 3 syllables to total number of words
		Count of acronyms
		Count of foreign words
		Count of words that occur twice (hapax dislegomena)
3	structural features	Count of sentences
		average sentence length in terms of words
		Average sentence length in terms of characters
		Count of sentences per paragraph
		Count of paragraphs
		Count of words per paragraph
4	readability Features	Flesch Kinkaid Grade Level
		Coleman Liau Index
		Automated Readability Index
		Gunning Fog Index
		LIX
		RIX Readability Index
		Flesch Reading Ease
		SMOG Index.

Table 3. Accuracies of author prediction when BOW approach is used with most frequent terms

Features/ Classification Algorithms	8000 Terms
NBM	61.65
SLOG	55.54
LOG	57.89
IBK	56.21
BAG	59.77
RF	63.32

By examining the dataset of different product reviews, it was observed that the differences in writing style of authors were identified in the terms used by the different authors. Based on this analysis, we selected 8000 most frequent terms for representing the document vector. In this work, the experimentation carried out with most frequent 8000 terms. It was observed that the obtained results were not satisfactory for author prediction when compared with existing approaches in Authorship Attribution.

Table 4 represents the accuracies of author prediction in BOW approach using various classifiers. In the BOW approach, unlike other classifiers, the RF classifier achieved an accuracy of 63.32% for author prediction when the most frequent 8000 terms were used to represent the document vector. It is not feasible to use all features extracted from the training documents for authorship attribution. The document representation with reduced set of features is a major requirement for most of the machine learning algorithms. The next section explains various feature selection measures proposed in different research domains to identify important features.

5. Feature selection algorithms and experimental results

Feature selection techniques used fewer amounts of features/tokens/words which are useful for identifying an author of a document easily, faster, and by using less computation power. Feature selection measures were used to identify a subset of features for describing the data, or in other words it is a method to reduce the high dimensionality by removing features which are not relevant for the classification. In this work, four feature selection measures such as information gain, chi-square, mutual information and NGL (Ng-Goh-Low) Coefficient were used in the experiment to find informative terms from 8000 terms.

5.1 Information gain (IG)

Information Gain (IG) selects features which reveal the most of the information about the classes [15]. Eq. (1) is used to compute Information gain of a term.

$$IG(t,c) = \sum_{c \in \{c_i, \bar{c_i}\}, t \in \{t_k, \bar{t_k}\}} P(t,c) \times \frac{P(t,c)}{P(t)P(c)} \quad (1)$$

Where, P(t,c) is the probability of the term t belongs to class c, P(t) is the probability of a term t, P(c) is the probability of a class c.

5.2 Chi-square

Chi-square feature selection measure computes the dependence between a term, t, and the class, c [16]. Chi-square computed using Eq. (2).

$$\chi^2 = \frac{N(AD-CB)^2}{(A+B)(A+C)(B+D)(C+D)} \quad (2)$$

Where, A, B is the number of documents in class c which contain the term t and which does not contain the term t respectively. C, D is the number of documents which does not belongs to class c and which contain the term t and which does not contain term t respectively. N is the total number of documents in the dataset.

5.3 Mutual information (MI)

Mutual information feature selection measure assumes that the term with higher class ratio is more efficient for classification [17]. Eq. (3) is used to compute Mutual information of a term.

$$MI = \log\left(\frac{A \times N}{(A+B)(A+C)}\right) \quad (3)$$

N, A, B, C parameters were defined in section 5.2.

5.4 NGL (Ng-Goh-Low) Coefficient

NGL correlation coefficient is a type of chi-square feature selection method [18].

$$NGL = \frac{\sqrt{N} \times (AD-CB)}{\sqrt{(A+B)(A+C)(B+D)(C+D)}} \quad (4)$$

Table 4. The accuracies of feature selection measures for author prediction

Features/ Classification Algorithms	IG	CHI	MI	NGL
NBM	63.81	68.84	64.61	69.65
SLOG	59.62	66.35	61.41	65.54
LOG	61.20	64.13	63.32	67.89
IBK	62.58	65.49	60.27	68.21
BAG	61.83	64.81	62.59	63.77
RF	65.17	69.87	66.31	71.19

The positive value of NGL measure represents the term t correlates with class c and negative value indicated the term t correlates with $\bar{c}$. The NGL method computed using Eq. (4). NGL assigns more weight to the terms which are having strong correlation with category c. N, A, B, C, D parameters were defined in section 5.2.

5.5 Experimental results of feature selection algorithms

The feature selection algorithms identified most informative terms in the 8000 terms. Top ranked 5000 terms were used from 8000 terms to represent the document vector. Different classification algorithms were used to generate the classification model. Table 5 shows the accuracies of author prediction when feature selection algorithms were used.

The NGL coefficient measure obtained good accuracy of 71.19% for author prediction when compared with other feature selection measures. The random forest classifier achieved better results compared to other machine learning classifiers. The results of feature selection measures for author prediction were not good when compared with existing approaches in Authorship Attribution. The next section explains the importance of term weight measures and discussed various term weight measures used in our experiment for author prediction.

6. Term weight measures

Term weight measures allocate suitable weight to the terms based on the information of terms distribution in the dataset. Traditional term weighting measures are Term Frequency (TF), binary and Term Frequency Inverse Document Frequency (TFIDF). Binary weight measure assigns 1 or 0 to the term based on the term presence or absence in a document. TF measure computes the frequency of a term in a document. TF may assign large weights to the common terms (a, an, the, of,

etc.,) which are weak in text discrimination. To overcome this shortcoming, TFIDF measure is proposed by the researchers to reduce the weight of common terms. In TFIDF measure, the IDF allocate more weight to the terms that were appeared in less number of documents. Although the TFIDF was proved in Information Retrieval domain and several text mining tasks for quantifying the term weights, but it is not most effective for Author Prediction because TFIDF disregard the class label information of the training documents. Therefore researchers are looking for alternative effective term weight measures in Authorship Attribution.

Based on the utilization of the class label information the term weight measures were categorized into two types such as unsupervised and supervised term weight measures. An unsupervised term weight measure does not use information regarding class label. The supervised term weight measure use class label information. In this work, 3 unsupervised and 5 supervised term weight measures were used to compute the weights of the terms.

6.1 Unsupervised term weight measures

6.1.1. TFIDF (Term Frequency Inverse Document Frequency)

The TFIDF measure [19] computed using Eq. (5).

$$TFIDF(t_i, d_k) = tf(t_i, d_k) \times \log\left(\frac{|N|}{DF_i}\right) \quad (5)$$

Where, $tf(t_i, d_k)$ is the number of times t_i occurred in document d_k, N is the number of documents in the dataset, DF_i is the number of documents in the dataset which contain the term t_i.

6.1.2. NDTW (Nonuniform Distributed Term Weight) Measure

The NDTW measure assigns more weight to the terms which are distributed non uniformly across the documents [14]. Eq. (6) shows the NDTW measure.

$$W_{t_{ij}} = w(t_i, p_j) = \log(TOTF_{ti}) - \sum_{k=1}^{m}\left(\frac{tf(t_i, d_k)}{TOTF_{ti}} \log\left[\frac{1+tf(t_i, d_k)}{1+TOTF_{ti}}\right]\right) \quad (6)$$

Where, $TOTF_{ti}$ is the total occurrence of term t_i in profile group p_j, $tf(t_i, d_k)$ is the frequency of term t_i in document d_k.

6.1.3. NDLTW (Normalized Document Length Term Weight) measure

A NDLTW Measure was proposed in [10] to avoid the differentiation of small sized and large sized documents. Eq. (7) represents the NDLTW measure.

$$W(t_i, p_j) = \frac{(1+\log(TF_i)) / (1+\log(AVGTF_i))}{\sum_{k=1}^{m}(1-slope) \times AVGUT_k + slope \times UT_k} \quad (7)$$

Where, TF_i is the number of times term t_i occurred in profile pj, AVGTFi is the ratio of TF_i to total number of terms in profile P_j, slope =0.2, UT_k number of unique terms in document d_k, $AVGUT_k$ is the ratio of UT_k to total number of terms in document dk.

6.2 Supervised term weight measures

6.2.1. RFTW (Relevance Frequency based Term Weight) measure

RFTW measure assigns more discriminative power to the terms which are discussed more in positive documents when compared with negative documents [20]. The RFTW measure is represented in Eq. (8).

$$tf * rf = tf \times \log\left(2 + \frac{A}{\max(1, C)}\right) \quad (8)$$

A, C parameters were defined in section 5.2.

6.2.2. Discriminative feature selection term weight (DFSTW) measure

DFS measure allocate more weight to the terms that are having high average term frequency in class c_j and the terms with high occurrence rate in most of the documents of c_j [21]. The DFSTW measure is showed in Eq. (9).

$$W(t_i, c_j) = \frac{tf(t_i, c_j) / df(t_i, c_j)}{tf(t_i, \bar{c_j}) / df(t_i, \bar{c_j})} \times \frac{A}{(A+B)} \times \frac{A}{(A+C)} \times \left|\frac{A}{(A+B)} - \frac{C}{(C+D)}\right| \quad (9)$$

Where, $tf(ti,c_j)$ is the term frequency of term t_i in class c_j, $df(t_i, c_j)$ is the number of documents contain the term t_i in class c_j and A, B, C, D parameters were defined in section 5.2.

6.2.3 TF-Prob measure

TF-Prob is a probability-based weight measure defined in [22]. The TF-Prob measure finds the weight of term t_k with respect to c_j is shown in Eq. (10).

$$w(t_k, c_j) = tf_k \times \log\left(1 + \frac{A}{B}\frac{A}{C}\right) \quad (10)$$

Where, tf_k is the frequency of term t_k in class c_j. A, B, C parameters were defined in section 5.2.

6.2.4. ICF-based term weighting schemes TF-IDF-ICSDF

Inverse Class Frequency (ICF) is similar to IDF in TFIDF, which is defined as the ratio of the total classes to the number of classes which contains the term. TF-IDF-ICSDF measure was proposed in [23]. The TF-IDF-ICSDF weight is computed by Eq. (11).

$$w(t_k) = tf_k \times \left(1 + \log\frac{N}{df_k}\right) \times \left(1 + \log\frac{m}{\sum_{j=1}^{m}\frac{df_{kj}}{N_j}}\right) \quad (11)$$

Where, N is the number of documents in the corpus, df_k is the number of documents contains the term t_k, m is the number of classes, and df_{kj} is the number of documents in class j contains the term tk, N_j is the number of documents in class c_j.(7)

6.2.5. SUTW measure

Supervised Unique Term Weight (SUTW) measure [24] as in Eq. (12) combines inner-document distribution, inter-class distribution and intra-class distribution information of terms to measure the weight of a term.

In Eq. (12), d_{tk} is the number of terms in document d_k, UT_k, $AVGUT_k$ was defined in section 6.1.3, A, B, C, D was defined in section 5.2.

$$Wt_{ij} = W(t_i, p_j) = \sum_{k=1, d_k \in p_j}^{m}\left(\frac{tf(t_i, d_k)}{tf(t_i, p_j)}\left[\frac{\log(d_k)}{0.8 \times AVGUT_k + 0.2 \times UT_k}\right]\right) \times \frac{A}{(A+B)} \times \frac{C}{(C+D)} \quad (12)$$

Table 5. The accuracies of author prediction for various term weight measures

Classifier/Term Weight Measures	Naïve Bayes Multinomial	Random Forest
TFIDF	66.29	69.45
NDTW	71.48	73.01
NDLTW	74.31	77.13
RFTW	78.39	82.92
TF-Prob	80.16	85.83
DFSTW	83.67	87.02
TF-IDF-ICSDF	85.71	89.11
SUTW	89.23	91.82

6.3 Experimental results of term weight measures

In this work, 8000 most frequent terms were extracted from the corpus to represent the document vectors. The BOW model is used to represent the document vectors with these 8000 terms. Various term weight measures were used to assign the weight to the terms in document vectors. Table 6 represents the accuracies of author prediction when different term weight measures were used to define the weight of the term. In table 6, the SUTW measure obtained highest accuracy of 91.82% for author identification when compared with all other term weight measures. It was identified that the supervised term weight measures achieved best accuracies for author prediction when contrasted with accuracies of unsupervised term weight measures. It was also noted that, the Random Forest classifier achieved good accuracies for most of the term weight measures when compared with other Classifier.

The accuracies of term weight measures are good when compared with most of the existing approaches for Authorship attribution. In this work, a new approach proposed to increase the accuracy of author prediction. The next section describes the proposed approach.

7. Document weight specific to author (DWA) approach

Fig. 2 shows the model for proposed Document Weight specific to Author (DWA) approach. The Authorship Attribution problems categorized in to two classes such as classical Authorship Attribution and social media Authorship Attribution. The classical Authorship Attribution problem concentrated on formally written documents such as newspapers, articles and books, while social media Authorship Attribution task concentrates on informal documents such as reviews, tweets and blogs.

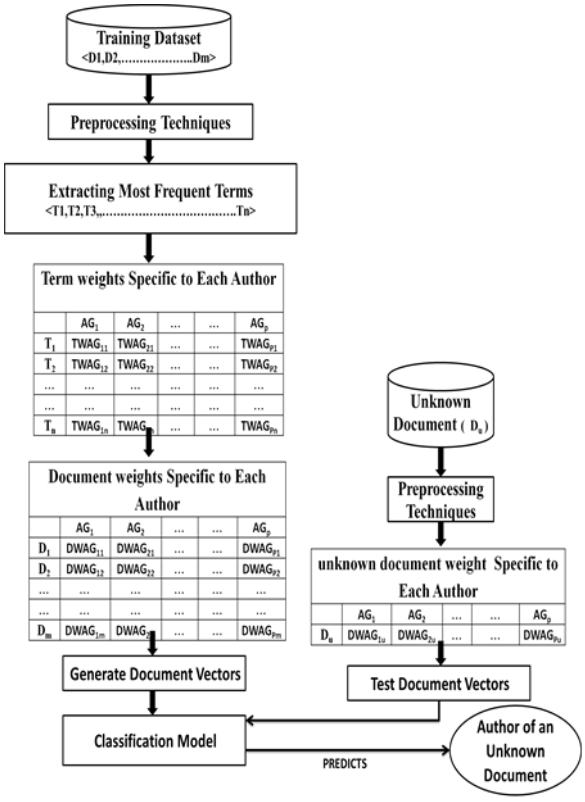

Figure.2 The model of proposed approach

Various researchers extracted different types of stylistic features ranging from character based features to syntactic features for classical Authorship Attribution. For social media Authorship Attribution it was observed that some of these features utilization was impractical because most of the documents contain more number of non-dictionary terms. For example, extraction of syntactic information from informal documents was very difficult. Therefore, most of the researchers proposed combinations of content and stylistic features for social media Authorship Attribution. In this work, content based features such as most frequent terms are used to differentiate the writing styles of the authors.

In general, every term is having a specific importance in different author groups. Different authors used a set of terms frequently in their writings. It is difficult to say whether the document was written by the particular author or not by considering some of the terms individually, but it is easy to predict the author by considering the combination of terms in the document. In BOW model, the document vectors were represented with the terms independently and the relationship between the terms were not captured. In the proposed DWA approach, the document vectors were represented with the document weights. The document weights were computed by combining the terms in that document and also consider the relationship of the terms. In this approach, it was also identified the best informative terms that are useful for differentiating the writing styles of the authors by using an efficient term weight measure. This was the main reason for obtaining best accuracy for predicting the author in authorship attribution.

In this model, first preprocessing techniques such as stopword removal and stemming were applied on the training dataset for preparing the data for effective features extraction. Extract most frequent terms from the updated dataset. Term weight measure is used to compute the weights of these extracted terms specific to author group. The dataset contains 10 author groups and each author group contains 400 documents. Document weight measure is used to compute the weight of the document by using the weights of the terms. The document vector is represented with these document weights and classification algorithms generates classification model by using these document vectors.

In this model, (D_1, D_2 ,……, D_m) is the set of documents in the dataset, (T_1, T_2, ……, T_n) is the set of most frequent terms, $TWAG_{pn}$ is the weight of the term T_n in the author group AG_p, $DWAG_{pm}$ is the weight of the document D_m in the author group AG_p. In this model, the term weight measure and document weight measure play an important role to improve the accuracy of author prediction. In this work, a new term weight measure is proposed to compute the weight of the term specific to author group of documents. The next subsection explains the proposed term weight measure.

7.1 Term weight easure

Various researchers proposed different types of term weight measures in different research areas. In this work, a new supervised term weight measure is proposed to compute the weights of the terms. The proposed supervised term weight measure is represented in Eq. (13). The main principle of this term weight measure is it assigns more weight to the terms which are having more frequency in interested author group and contained in more number of documents in interested author group.

In this measure, $tf(t_i,d_k)$ is the term frequency in document d_k, DF_k is the total number of terms in a document d_k.

$\sum_{x=1, d_x \in AG_p}^{m} tf(t_i, d_x)$ gives the total count of the term t_i in all the documents of author group AG_p.

$$W\left(t_i, d_k \in AG_p\right) = \frac{tf(t_i, d_k)}{DF_k} \times \frac{\sum_{x=1, d_x \in AG_p}^{m} tf(t_i, d_x)}{1 + \left(\sum_{y=1, d_y \notin AG_p}^{n} tf(t_i, d_y) \right)} \times \frac{\sum_{x=1, d_x \in AG_p}^{m} DC(t_i, d_x)}{1 + \left(\sum_{y=1, d_y \notin AG_p}^{n} DC(t_i, d_y) \right)} \quad (13)$$

$\sum_{y=1, d_y \notin AG_p}^{n} tf(t_i, d_y)$ Gives the total count of the term t_i in all the documents of all author groups except AG_p

$\sum_{x=1, d_x \in AG_p}^{m} DC(t_i, d_x)$ Gives the number of documents in author group AG_p contains the term t_i

$\sum_{y=1, d_y \notin AG_p}^{n} DC(t_i, d_y)$ Gives the number of documents in all author groups except AG_p contains the term t_i

7.2 Document weight measure

In this work, a document weight measure is used proposed by Raghunadha reddy et al., [24]. The document weight measure determines the weight of a document by considering different information of terms in a document. Eq. (14) represents the document weight measure used in our experiment.

$$W\left(d_k, AG_p\right) = \sum_{t_i \in d_k, d_k \in AG_p} TFIDF(t_i, d_k) \times W\left(t_i, AG_p\right) \quad (14)$$

This measure used two types of information of terms such as TFIDF (Term Frequency and Inverse Document Frequency measure) weight of a term and the term weight calculated by term weight measure to compute the weight of a document. In this measure, w (d_k, AG_p) is the weight of document d_k in author group AG_p.

7.3 Experimental results of proposed DWA approach

Table 7 shows the DWA approach accuracies for author prediction.

Table 7. The Accuracies of DWA approach for author prediction

Features/ Classification Algorithms	**8000 Terms**
NBM	91.56
SLOG	85.17
LOG	89.63
IBK	87.29
BAG	91.45
RF	95.89

The experimentation carried out with 8000 most frequent terms for generating classification model. It was observed that the obtained results were best for author prediction when compared with most of the existing approaches [6, 8, 9, 11, 13] for author prediction in Authorship Attribution. When compared with all classifiers the Random Forest classifier achieved highest accuracy of 95.89% for author prediction.

8. Conclusion and future scope

In this work, the experimentation is carried out with stylistic features, most frequent terms and feature selection measures with BOW model and proposed DWA model. The proposed model achieved an accuracy of 95.89% for author prediction when Random Forest classifier was used. The BOW approach with feature selection measures obtained an accuracy of 71.19% for author prediction when Random Forest classifier is used. In BOW approach the terms are independently participated in the classification process, but in proposed DWA model the terms are collaboratively in the form of document weight participated in the classification process. This is the main reason for obtaining good accuracies in the proposed model.

In our future work, it is planned to consider the domain characteristics and categorical features while computing a document weight. It is also planned to usage of semantic and syntactic structure of the language while assigning weights to the document.

References

[1] J. Schler, M. Koppel, S. Argamon, and J. Pennebaker, “Effects of Age and Gender on Blogging”, In: *Proc. of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, 2006.

[2] E. Stamatatos, “A Survey of Modern Authorship Attribution Methods”, *Journal of the American Society for Information Science*

and Technology, Vol.60, No.3, pp.538-556, 2009.

[3] M. Koppel, J. Schler, and E. Bonchek-Dokow, "Measuring differentiability: Unmasking pseudonymous authors", *The Journal of Machine Learning Research*, Vol.8, pp.1261–1276, 2007.

[4] M. Koppel, S. Argamon, and A. Shimoni, "Automatically categorizing written texts by author gender", *Literary and Linguistic Computing*, Vol.17, No.4, pp.401-412, 2002.

[5] P. Juola, "Authorship Attribution", *Foundations and Trends in Information Retrieval*, Vol.1, No.3, pp.233-334, 2008.

[6] L. Tanguy, F. Sajous, B. Calderone, and N. Hathout, "Authorship attribution: using rich linguistic features when training data is scarce", In: *Proc. of CLEF 2012 Evaluation Labs and Workshop*, 2012.

[7] J. Kapociute-Dzikiene, A. Utka, and L. Sarkute, "Authorship Attribution of Internet Comments with Thousand Candidate Authors", *Information and Software Technologies. Communications in Computer and Information Science*, Vol.538, pp 433-448, 2015

[8] S. Ruseti and T. Rebedea, "Authorship Identification Using a Reduced Set of Linguistic Features", In: *Proc of CLEF 2012 Evaluation Labs and Workshop*, 2012.

[9] G. K. Mikros and K. Perifanos, "Authorship identification in large email collections: Experiments using features that belong to different linguistic levels", In: *Proc. of CLEF 2011 Evaluation Labs and Workshop*, 2011.

[10] A. Singhal, C. Buckley, and M. Mitra, "Pivoted document length normalization", In: *Proc. of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 21-29, 1996.

[11] A. Rexha, S. Klampfl, M. Kröll, and R. Kern, "Towards authorship attribution for bibliometrics using stylometric features", In: *Proc. of CEUR Workshop*, pp. 44-49, 2015.

[12] S. Raghavan, A. Kovashka, and R. Mooney, "Authorship Attribution Using Probabilistic Context-Free Grammars", In: *Proc. of the ACL Conference Short Papers*, pp.1-3 2010.

[13] N. Akiva, "Authorship and Plagiarism Detection Using Binary BOW Features", In: *Proc. of CLEF 2012 Evaluation Labs and Workshop*, 2012.

[14] S.F. Dennis, "The Design and Testing of a Fully Automated Indexing-Searching System for Documents Consisting of Expository Text", *Informational Retrieval: A Critical review*, pp.67-94, 1967

[15] R. Mukras, N. Wiratunga, R. Lothian, S. Chakraborti, and D. Harper, "Information gain feature selection for ordinal text classification using probability re-distribution", In: *Proc. of IJCAI Textlink Workshop*, pp.1-10, 2012.

[16] M. Z. F. Thabtah, M. Ali, H. Eljinini, and W. M. Hadi, "Nave bayesian based on chi square to categorize arabic data", *Communications of the IBIMA*, Vol.10, No. 20, pp.158-163, 2009.

[17] S. Li, R. Xia, C. Zong, and C.-R. Huang, "A framework of feature selection methods for text categorization", In: *Proc. of ACL/AFNLP'09*, pp. 692-700, 2009.

[18] G. Uchyigit and M. Ma, "Personalization techniques and recommender systems", *Machine Perception and Artificial Intelligence*, World Scientific, Vol. 70, 2008.

[19] G. Salton and C. Buckley, "Term weighting approaches in automatic text retrieval", *Information Processing & Management*, Vol.24, No.5, pp.513-523, 1988.

[20] M. Lan, C.L. Tan, J. Su, and Y. Lu, "Supervised and traditional term weighting methods for automatic text categorization", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.31, No.4, pp.721–735, 2009.

[21] W. Zong, F. Wu, L. Chu, and D. Sculli, "A discriminative and semantic feature selection method for text categorization", *International Journal of Production Economics*, pp. 215-222, 2015.

[22] Y. Liu, H. T. Loh, and A. Sun, "Imbalanced text classification: A term weighting approach", *Expert Systems with Applications*, Vol.36, No.1, pp. 690–701, 2009.

[23] F. Ren and M.G. Sohrab, "Class-indexing based term weighting for automatic text classification", *Information Sciences*, Vol.236, pp.109–125, 2013.

[24] T. Raghunadha Reddy, B. Vishnu Vardhan, and P. Vijayapal Reddy, "Profile specific Document Weighted approach using a New Term Weighting Measure for Author Profiling", *International Journal of Intelligent Engineering and Systems*, Vol.9, No.4, pp. 136-146, 2016.