INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Profile specific Document Weighted approach using a New Term Weighting Measure for Author Profiling

T. Raghunadha Reddy[1]* **B. Vishnu Vardhan[2]** **P. Vijayapal Reddy[3]**

[1]*Vardhaman College of Engineering, Hyderabad, Telangana, India*
[2]*JNTUH College of Engineering, Jagtiyal, Telangana, India*
[3]*Matrusri Engineering College, Saidabad, Hyderabad, Telangana, India*
** raghu.sas@gmail.com*

Abstract: Author Profiling is a text classification technique to predict the demographic features like age, gender, native language, location, educational background of the authors by analyzing their writing styles. Term weight measures identify the term discriminators for classifying the documents by assigning suitable weights to the terms. In this work, a supervised unique term weight measure is proposed to measure the significance of each term in the document. The proposed term weight measure is compared with four benchmark term weight measures such as *TF*, *TFIDF*, *tf * rf*, *iqf * qf * icf*. The experimental results show that the proposed term weight measure achieved the best performance among all term weight measures. The existing models fail to capture the relationship between terms and documents. To overcome the problem of independence among the terms within the document, in this work a new model has proposed by using second order representations between documents and profiles. In the second order representation, initially the relation between the terms within the document has established then, recognize the relationship among the documents and profiles. The performance of the proposed model is compared with existing model using various classifiers on reviews corpus. The results shows that the proposed approach with new term weight measure out performs for predicting gender, age and location of the authors.

Keywords: TF; TFIDF; tf * rf; iqf * qf * icf; Term Weight Measure; BOW model; PDW model; Gender Prediction; Age Prediction; Location Prediction; Author Profiling

1. Introduction

In recent years the information is growing rapidly on the web, especially on blogs, twitter, reviews and social networks. Most of the text in the web is anonymous. The users are not interested to specify their correct profiles while uploading their data on to the web. Automatic identification of the profiles of this anonymous text is a difficult problem. Author Profiling is an interesting research area attracted by several researchers to speculate the details of the anonymous text in the web.

Author Profiling is an important technique in the present information era which has applications in forensic analysis, security, marketing and educational background. Forensics is a field to analyze signatures, documents, and anonymous letters to identify the perpetrator of a crime. Author Profiling helps in forensic analysis and security by analyzing the characteristics of writing styles. In the marketing domain, the consumers were provided with a space to review the product. Most of the reviewers are not comfortable in revealing their personal identity. In this case, these anonymous reviews were analyzed by using Author Profiling techniques to classify the consumers based on their age, gender, occupation, native language and country. Based on the classification results, companies try to adopt new business strategies to serve the customers. Author Profiling is also beneficial in educational domain by analyzing a large set of pupil's text. It helps in revealing the

exceptional talent of the students and also helps in estimating the suitable level of knowledge of each student or a student group in the educational forum.

In Author Profiling, the profiles of the authors are identified based on the writing styles of the authors. According to Koppel et al. [2], men use more number of determiners and quantifiers and woman use more number of pronouns than men in their writings. Similarly the male authors stress more on topics related to sports, politics and technology whereas the female authors write about topics like beauty, kitty parties and shopping. Generally the writing styles of the authors vary based on the selection of topics and the writing styles like choice of words and grammar rules. In an observation [4], the female write more about wedding styles and males write more about technology and politics. Further female use more adjectives and adverbs than male authors. The users in age group of 13-17 describe the topics related to adolescence, school activities and immature crush, the users from 23-27 age group write more about pre-marital affairs, favourite heroines/heroes and college life and the users belonging to 33-47 age group post more about post-marriage life and corporate/social activities [3].

The main focus of this paper is to predict the profiles of the authors like gender, age and location in reviews domain. This paper is organized in seven sections. The related work in Author Profiling is described in section 2. The dataset characteristics and measures used to evaluate the model are explained in section 3. Section 4 describes the existing term weight measures and proposed term weight measure. The proposed model is explained in section 5. Section 6 discusses the accuracies of author profiles prediction using Naïve Bayes Multinomial and Random Forest algorithms. Section 7 concludes this work with future possible extensions to the proposed work.

2. Related work

Text classification uses a set of features and machine learning techniques to assign predefined classes to text documents. In the procedure of text classification, the raw text documents are to be converted in to a vector representation since the classifiers never process text documents directly. Traditional approaches to Author Profiling used the feature frequency to specify the importance of a feature in a document. The researchers realized that the feature frequency is not sufficient to find the importance of a feature. Then, TFIDF (Term Frequency Inverse Document Frequency) measure was proposed to determine the weight of a feature based on the feature frequency and the number of documents contains the feature in a corpus [1]. Later, the researchers proposed various weight measures to compute the weight of the features. The number of features, its weight measures and machine learning algorithms influence the prediction accuracies of the author profiles in Author Profiling.

Juan Soler Company et al. experimented [5] their work on the corpus of New York Times opinion blogs. They tried with different combinations of features including word based, character based, sentence based, dictionary based and syntactic features for gender prediction and achieved good accuracy when all the features were combined. It was observed that the accuracy was reduced when Bag of Words approach is applied with 3000 words having most TFIDF values.

Maria De-Arteaga et al. [6] tried with lexical, stylistic and corpus statistic features, Corpus statistic features including unsupervised corpus statistics such as IR features (IDF and TFIDF), entropy measures, KL divergence measure and cross entropy measure and supervised corpus statistics including gender score measure, bayes score, supervised KL-divergence, supervised cross entropy, supervised lexicon extraction features. In their observation, the supervised corpus statistics were best predictors for age prediction compared to unsupervised corpus statistics and also observed that the lexical and stylistic features are more suitable for age prediction than gender prediction.

Wee-Yong Lim et al., used [7] TFIDF scores of words to find the rare or common words in the entire corpus. Seifeddine Mechti et al., computed [8] the ranked list of words, then group these words into classes according to their similarity. The TFIDF measure was used to calculate the scores of each class for each document to find the stylistic differences between male and female. Suraj Maharjan et al., recognized [9] word n-grams as features and TFIDF as the weighting measure. TFIDF scores of the word n-grams were used to filter the n-grams that were not been used by at least two authors. Andreas Grivas et al., experimented [9] with TFIDF scores of word n-grams and bag of words to generate feature vectors.

Alonso Palomino-Garibay et al., tested [11] on tweets corpus and represented each tweet with a bag of words in a vector space. TFIDF measure was used to assign a value to each word in a vector. Octavia-Maria S et al., used [12] the combination of type/token ratio and TFIDF scores of character n-grams. The TFIDF scores were extracted from scikit-learn's TfidfVectorizer (). It was observed

that this combination of features obtained good accuracies for Dutch and Spanish language and also observed that the best TFIDF scores were obtained for character level n-grams in the range 'n' value between 2 to 6.

Several researchers proposed various combinations of features like lexical, character based, syntactic, structural, semantic and readability features to differentiate the writing styles of the authors [13]. Estival D. et al. collected [14] 9836 emails and extracted 689 features of character level, lexical and structural features. Various machine learning algorithms such as J48, RandomForest, IBK, JRip, SMO, libSVM, Bagging, AdaBoostM1 were applied on the corpus. It was observed that SMO obtained best accuracy for gender prediction compared to other classifiers.

Dang Duc, P. et al. experimented [15] with 3524 pages of 73 Vietnamese bloggers. 298 features including word based and character based features were extracted from the blogs corpus. It was observed that the word based features contributed more to gender prediction than character based features and the classifier IBK obtained a good accuracy for gender prediction using combination of word and character based features. In another work [16], the researchers collected 1000 blog posts of 20 bloggers from Greek language and extracted standard stylometric features and 300 most frequent n-grams such as word n-grams and character n-grams. In their observation, longer sequences of word n-grams and character n-grams increase the accuracy for gender prediction using Support Vector Machine.

In Koppel, M. et al. [2], 566 documents were taken from the British National Corpus (BNC). They achieved an accuracy for gender prediction using 1081 features. Argamon, S. et al. [17] collected corpus of blog posts of 19320 blog authors. The result accuracy is achieved for gender dimension by using both content based and stylistic features. It was observed that the style based features were most useful to discriminate the gender. In another work [3], they achieved better accuracy using 1502 features of content based and stylistic features on 37478 blogs.

Man lan et al., [18] proposed a new supervised term weight measure tf * rf (term frequency * relevance frequency) measure for text categorization. They compared the performance of this measure with various term weight measures like binary representation, tf, tf.idf (term frequency. inverse document frequency), tf.chi2 (term frequency.Chi square), tf.ig (term frequency. information gain), tf.logOR (term frequency . log OddRatio) and observed that the proposed term weight measure show best performance for categorization in various datasets such as Reuters news corpus, 20 newsgroups corpus and ohsumed corpus.

Xiaojun Quan et al., proposed three supervised term weight measures namely qf * icf, iqf * qf * icf and vrf for text categorization. They compared these measures with existing measures such as tf, tf*idf, tf*ig, tf*chi2, tf*OR and tf*rf and it was observed that iqf * qf * icf achieved the best performance among all the term weight measures.

The existing approaches to Author Profiling suffered from the high dimensionality problem and fail to capture the relationship between the features. In this work, a Profile specific Document Weight (PDW) approach is proposed to address the problems faced in existing approaches of Author Profiling.

3. Dataset and Evaluation Measure

3.1 Dataset characteristics

The corpus used in this work was collected from hotel reviews website www.TripAdvisor.com, and it contains 4000 English reviews about different hotels. The corpus was constructed carefully to ensure its quality with regard to text cleanliness and annotation accuracy. In order to make this dataset applicable to Author Profiling and to ensure its quality, reviews containing less than 5 lines of text were excluded from our dataset and the reviews written by the authors whose gender was given in their user profile. In this work, three author profiles such as gender, age and location were considered for analysis. The corpus is balanced in gender and location dimension, but unbalanced in case of age dimension. Table 1 depicts the characteristics of the reviews dataset for gender and age dimension. The characteristics of the reviews dataset for location dimension are represented in Table 2.

Table 1. Dataset characteristics of gender and age profiles

S no	Age Group	Number of Documents	Number of Male Documents	Number of Female Documents
1	18-24	400	200	200
2	25-34	1000	500	500
3	35-49	1000	500	500
4	50-64	1000	500	500
5	65_and	600	300	300
Total		4000	2000	2000

Table 2. Dataset characteristics of location profile

S No	Country	Number of Documents
1	Australia	400
2	Brazil	400
3	China	400
4	Germany	400
5	India	400
6	Japan	400
7	Pakistan	400
8	Russia	400
9	UK	400
10	USA	400

The profile groups for gender profile are male and female, for age profile, the profile groups are 18-24, 25-34, 35-49, 50-64, 65_and_above and for location profile, Australia, Brazil, China, Germany, India, Japan, Pakistan, Russia, UK and USA are the profile groups.

3.2 Evaluation Measures

The existing approaches to Author Profiling used various measures such as precision, recall, F1-score and accuracy for evaluating their system performance. In this work, Accuracy measure is used to measure the performance. Accuracy is the ratio of number of documents correctly predicted their profiles to the total number of documents considered for evaluating the classifier.

$$Accuracy = \frac{Number\ of\ \text{documents correctly predicted their profile}}{Total\ number\ of\ test\ documents}$$

4. Term Weight Measures

In this work, various term weight measures such as TF, TFIDF, *tf * rf* measure (**TWM-I**), iqf * qf * icf Weight Measure (**TWM-II**) and proposed Supervised Unique Term Weigh Measure (**TWM-III**) are investigated to compute a term weights to represent the document vectors for generating classification model.

4.1 Term Frequency (TF)

Term frequency is the number of times the term occurred in a document.

4.2 Term Frequency Inverse Document Frequency (TFIDF)

Term Frequency Inverse Document Frequency (TFIDF) measure was proposed by Jones [1], which is used to calculate the weight of the term in a particular document. TFIDF measure as in equation (1) assigns weight to a term based on the term frequency and number of documents contains the term in a corpus of documents.

$$TFIDF(t_i, d_k) = tf(t_i, d_k) * \log\left(\frac{|D|}{|1 + DF_{ti}|}\right) \quad (1)$$

Where, $|D|$ is the total number of documents in the corpus, DF_{ti} is the number of documents contains the term t_i in the corpus of documents.

4.3 *tf * rf* Measure (TWM-I)

Term weight measures assigns a suitable weights to the terms based on their importance in the text. This term weight measure [18] was taken from the text categorization domain. This measure considers the appearance of term *t* in positive documents and negative documents. The basic idea of this measure is the terms which are having high frequency in positive category documents having more discriminative power to select positive samples than negative samples. The *tf * rf* measure is shown in equation (2).

$$tf * rf = tf * \log\left(2 + \frac{a}{\max(1, c)}\right) \quad (2)$$

Where, *a* is the number of documents in positive category that contain the term t_i, *c* is the number of documents in the negative category that contain the term t_i.

4.4 *iqf * qf * icf* Measure (TWM-II)

This weight measure as shown in equation (4) was proposed in [19]. *qf* (question frequency) of term t_i is the number of documents contains the term t_i in the interested category of documents. icf (inverse category frequency) computes the discriminative power of a term in all the categories. iqf (inverse question frequency) similar to idf (inverse document frequency) which computes the discriminative power of term t_i over all the documents.

$$iqf * qf * icf = \log\left(\frac{N}{tp + fn}\right) * \log(tp + 1) * \log\left(\frac{|C|}{cf} + 1\right) \quad (3)$$

Where, N is the number of documents in the dataset, tp is the number of positive documents that contain term t_i, fn is the number of negative documents contains the term t_i, $|C|$ is the number of categories, cf (category frequency) of term t_i is the number of categories in which t_i occurs.

4.5 Proposed Measure: Supervised Unique Term Weight Measure (TWM-III)

Supervised Unique Term Weigh Measure is proposed in this work. TF, TFIDF are the unsupervised term weighting measures. TWM-I, TWM-II, TWM-III are supervised term weighting measure, which uses known membership information of documents.

$$Wt_{ij} = W(t_i, p_j) = \sum_{k=1, d_k \in p_j}^{m} \left(\frac{tf(t_i, d_k)}{tf(t_i, p_j)} \left[\frac{\log(d_k)}{0.8 * AVGUT_k + 0.2 * UT_k} \right] \right) \times \frac{a_{ij}}{(a_{ij} + b_{ij})} \times \frac{c_{ij}}{(c_{ij} + d_{ij})} \quad (4)$$

Where, $tf(t_i, d_k)$ is the term frequency of t_i in the document d_k, $tf(t_i, p_j)$ is the term frequency of t_i in the profile group p_j, d_{tk} is the number of terms in a document d_k, a_{ij} is the number of documents of profile group p_j that contain the term t_i, b_{ij} is the number of documents of profile group p_j that do not contain the term t_i, c_{ij} is the number of documents that contain the term t_i but do not belongs to profile group p_j, d_{ij} is the number of documents that do not contain the term t_i and do not belongs to profile group p_j.

The existing approaches used the relationship between the features and profile of a document. The document is a source to extract the information for features. As of our knowledge the relationship between document and profile is not yet exploited in Author Profiling. The next section explains proposed model, which establish a relationship between a document to profile and features to documents.

5. Profile specific Document Weighted (PDW) Model

Every term is having a specific importance in different profile groups. For example 'bowl' is a word that is occurred in male documents in the context of cricket and in female documents in the context of kitchenware. If a new document contains bowl word, the document is written by male or female is not predicted certainly. The proposed model represents the document with document weights not with the features in that document. In this model, the weight of bowl word is computed in all the documents of male and female. Maintain the term weights separately specific to each profile group of gender. The document weights are calculated specific to each profile group by aggregating the weights of the terms specific to profile group. The architecture of proposed model is depicted in Fig. 1.

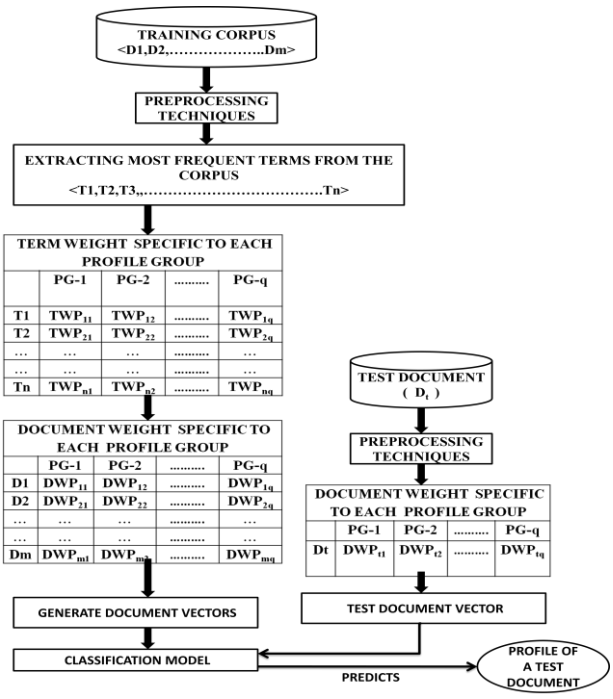


Figure.1 The architecture of PDW Model

In this model, $\{D_1, D_2, \ldots\ldots D_m\}$ is a list of documents in the corpus, $\{T_1, T_2, \ldots\ldots T_n\}$ is a list of vocabulary terms, $\{p_1, p_2, \ldots\ldots p_q\}$ is a set of profile groups in a profile P. TWP_{nq}, DWP_{mq} is the weight of the term T_n and D_m respectively in the profile group p_q.

The procedure in the proposed approach

1. Collect the corpus.
2. Apply preprocessing techniques for stop words removal and stemming is performed using porter stemming algorithm.
3. Extract most frequent terms.
4. In the first order representation, compute term weights for each profile group using term weight measures.
5. In the second order representation, document weights are determined for each profile group by aggregating the weights of the terms in a document.
6. Generate document vectors with document weights as the document vectors are used to build a classification model.

The profiles of an anonymous document are predicted using classification model. In this procedure, identification of suitable weight measures for calculating term weight and document weight is important. The term weight measures that are discussed in section 4 are used to compute the term weight. The following sub section 5.1 discuss about the proposed document weight measure.

5.1 Document Weight Measure

The proposed document weight measure as in equation (5) is used to calculate the weight of a document on corpus of each profile group. This measure used the combination of term weights that are specific to document and specific to profile group. The TFIDF measure used to compute term weights specific to a document and term weight measures are used to determine the term weights specific to profile. The document weight computation considers the correlation between the terms in that document. The document weight computation is expressed as below

$$W_{dkj} = \sum_{t_i \in d_k, d_k \in p_j} TFIDF(t_i, d_k) * W_{tij} \quad (5)$$

Where, W_{dkj} is the weight of document d_k in the profile p_j, Wt_{ij} is the weight of a term t_i in the corpus of profile group p_j.

The collections of training documents are finally represented using equation (6)

$$Z = \bigcup_{d_k \in p_j} (z_k, c_j) \quad (6)$$

Where, $z_k = \{W_{dk1}, W_{dk2,\ldots,}W_{dkq}\}$ and c_j is a class label of profile p_j. This representation of document addresses the problem of high dimensionality in existing approaches.

The vector Z contains weights of a document specific to each profile group with document profile label. This representation shows the correlation between the documents and their profiles.

6. Experimental Results

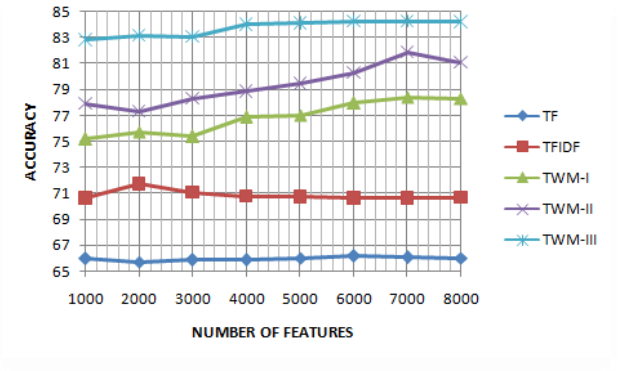
The BOW model is a simple representation of a document in text classification. In this model, the textual document is represented as the bag of its words (terms), keeping multiplicity but ignoring grammar and word order. The frequency of each word is used as a feature to train the classifier.

In this experiment, 8000 most frequent terms that are occurred at least two times in the total corpus are extracted. From the analysis of the reviews dataset, it was observed that 500 features are not sufficient to allow an effective discrimination amongst the text documents in different profile groups. In the following sections, the performances of the five term weight measures in BOW and PDW models are compared using Naive Bayes Multinomial and Random Forest algorithms.

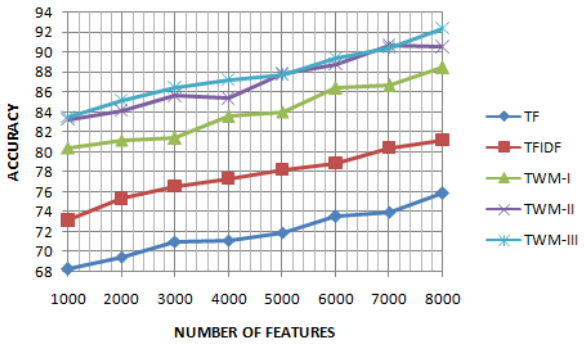
6.1 Gender Prediction

6.1.2. Accuracy of BOW and PDW models using Naive Bayes Multinomial algorithm for Gender Profile prediction

The performance of five term weight measures on BOW model and PDW model is depicted in Figure.2. When the number of features varies from 1000 to 8000, it was observed that the accuracy of proposed term weight measure is increased by good amount in PDW model when compared to BOW model. In PDW model, amongst the five weight measures TF and TFIDF perform relatively poorer than others when compared with BOW.



(a)



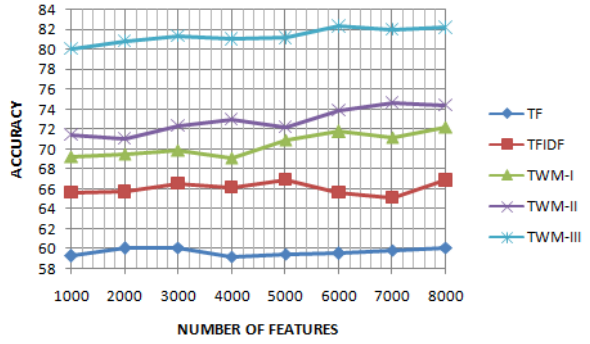
(b)

Figure.2 Performance of five term weight measures using Naïve Bayes Multinomial algorithm for gender prediction in (a) BOW Model (b) Proposed Model

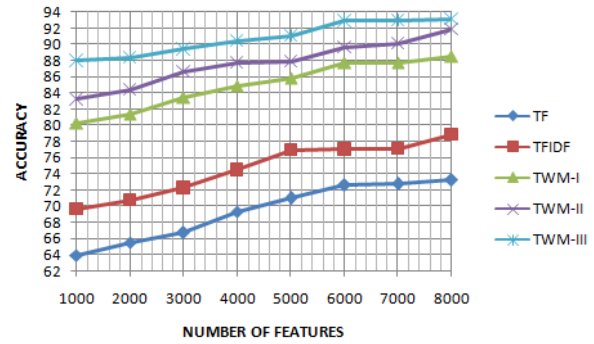
The TWM-III produces accuracies that are superior to the other four term weight measures over the full range of various numbers of features in BOW model, but where as in PDW model TWM-III accuracy is slightly falling between 6000 and 7000 range of features when refer to TWM-II. In PDW model the TWM-III accuracy is increased continuously when the number features varies, but there is a decrement in BOW model when the number of features changed from 2000 to 3000. All five term weight measures produce good accuracies in PDW model when compared to the accuracies of BOW. The proposed term weight measure generates an accuracy of 92.37% for gender prediction in PDW model, it is far better than the accuracy of 84.25% in BOW for gender prediction using Naïve Bayes Multinomial algorithm.

6.3.2. Accuracy of BOW and PDW models using Random Forest algorithm for Gender Profile

Figure.3 represents the accuracies of gender prediction in BOW model and PDW model when Random Forest classifier is used.



(a)



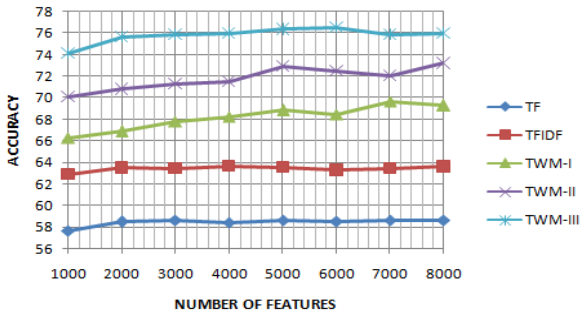
(b)

Figure.3 Performance of five term weight measures using Random Forest algorithm for gender prediction in (a) BOW Model (b) Proposed Model

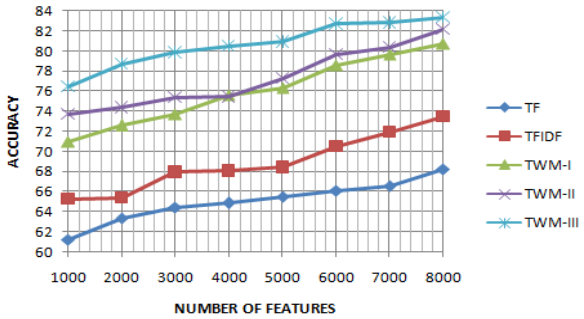
The proposed term weight measure (TWM-III) produced highest accuracy (93.15%) for gender prediction in PDW model when compared to BOW (82.30%). In PDW model, the TWM-III achieved highest accuracy when the number of features is 8000, but in BOW model most frequent 6000 terms achieved highest accuracy. Amongst the five term weight measures, TF and TFIDF perform relatively poorer than others in both BOW and PDW models. The proposed term weight measure shows accuracy reduction when the number of features changes from 3000 to 4000 in BOW model where as in PDW model the accuracy increased continuously over the full range of various numbers of features. In BOW model, the TWM-II produces higher accuracy of 74.67% than TWM-I (71.19%) when the number of features is 7000. In PDW model, The TWM-II produces higher accuracy of 91.89% than TWM-I (88.53%) when the number of features is 8000.

6.2 Age Prediction

6.2.1. Accuracy of BOW and PDW models using Naive Bayes Multinomial algorithm for Age Profile



(a)

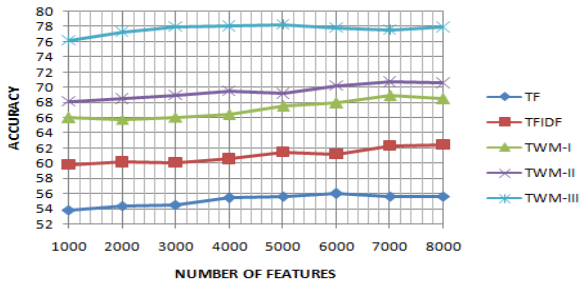


(b)

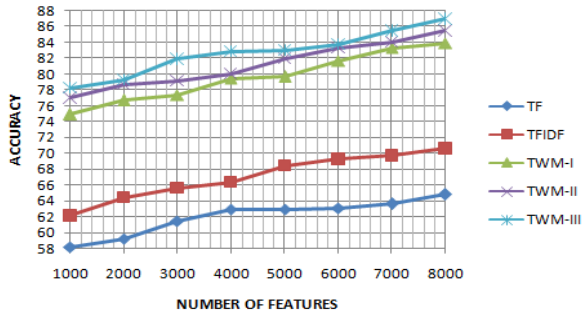
Figure.4 Performance of five term weight measures using Naïve Bayes Multinomial algorithm for age prediction in (a) BOW Model (b) Proposed Model

The accuracies of age prediction in BOW model and PDW model using five term weight measures are displayed in Figure.4. In PDW model, among the five weight measures the TF and TFIDF produce poor results than others. The TWM-II measure obtained good accuracies than TWM-I over the full range of features in BOW model, where as in PDW model TWM-I measure achieved good accuracies than TWM-II when the number of features 4000. The proposed term weight measure shows accuracy reduction when the number of features changes from 6000 to 7000 in BOW model, but the PDW model shows increase in accuracy when the number of features increases. The proposed term weight measure (TWM-III) produces accuracies that are superior to the other four term weight measures over the full range of various numbers of features in both models. The TWM-III produces a good accuracy of 83.31% for age prediction in PDW model which is far better than the accuracy (76.50%) in BOW model. In BOW model, TWM-II measure produces higher accuracy of 73.19% than TWM-I (69.27%) when the number of features is 8000.

6.2.2. Accuracy of BOW and PDW models using Random Forest algorithm for Age Profile



(a)



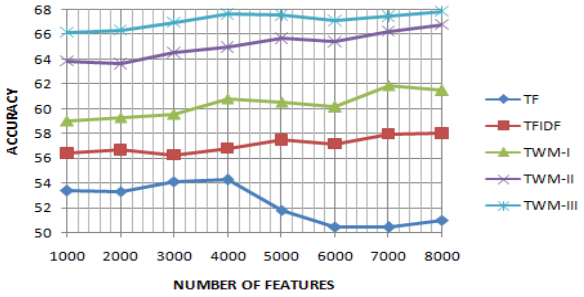
(b)

Figure.5 Performance of five term weight measures using Random Forest algorithm for age prediction in (a) BOW Model (b) Proposed Model

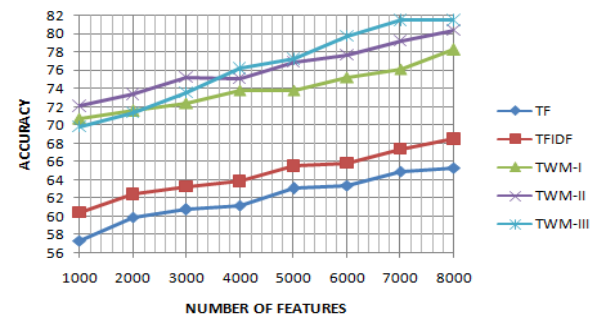
Figure.5 shows the accuracies of age prediction in BOW and PDW model using five term weight measures when Random Forest classifier is used. In PDW model the TWM-II measure obtained good accuracies than TWM-I over the full range of features. In PDW model, TWM-II produces higher accuracy of 85.48% than TWM-I (83.91%) when the number of features is 8000. In BOW model, The TWM-II produces higher accuracy of 70.69% than TWM-I (68.89%) when the number of features is 7000. The TWM-III produces a good accuracy of 86.97% for age prediction in PDW model which is far better than the accuracy (78.30%) in BOW model. The TWM-III shows accuracy reduction when the number of features changes from 5000 to 6000 in BOW model, where as in PDW model the accuracy consistently increases when the number of features increases. The TWM-III produces accuracies that are superior to the other four term weight measures over the full range of various numbers of features in both models.

6.3 Location Prediction

6.3.2. Accuracy of BOW and PDW models using Naive Bayes Multinomial algorithm for Location Profile



(a)



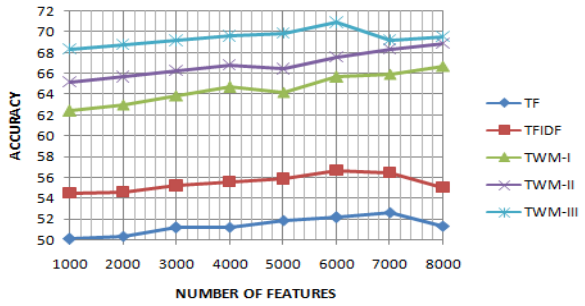
(b)

Figure.6 Performance of five term weight measures using Naïve Bayes Multinomial algorithm for location prediction in (a) BOW Model (b) Proposed Model

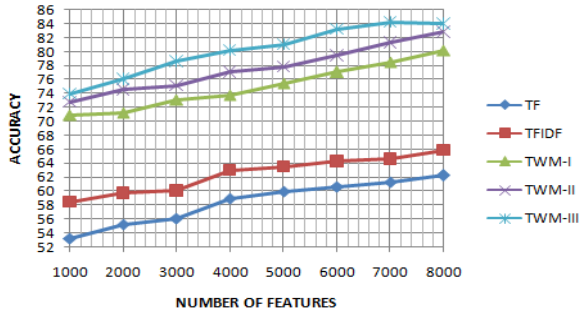
The accuracies of location prediction in BOW and PDW model using five weight measures are depicted in Figure.6. In PDW model, TWM-II produces higher accuracy of 80.36% than TWM-I (78.27%) when the number of features is 8000 where as in BOW model, for the same number of features the TWM-II produces higher accuracy of 66.79% than TWM-I (61.47%). In PDW model, the TMW-II produces higher accuracies than TWM-III when the number of features changed from 1000 to 3500 and TMW-I measure produce good accuracy than TWM-III when the number of features changed from 1000 to 2000. The proposed term weight measure (TWM-III) produces accuracies that are superior to the other four term weight measures over the full range of various numbers of features in BOW model. The TWM-III produces a good accuracy of 81.49% for location prediction in PDW model which is far better than the accuracy (67.85%) in BOW model.

6.3.2. Accuracy of BOW and PDW models using Random Forest algorithm for Location Profile

Figure.7 depicts the accuracies of location prediction in BOW and PDW model using five term weight measures.



(a)



(b)

Figure.7 Performance of five term weight measures using Random Forest algorithm for location prediction in (a) BOW Model (b) Proposed Model

TF and TFIDF measures produce poor results among the five weight measures in both BOW and PDW model. In BOW model, TWM-II produces higher accuracy of 68.91% than TWM-I (66.72%) when the number of features is 8000 where as in PDW model, TWM-II produces higher accuracy of 82.79% than TWM-I (80.21%) when the number of features is 8000. The TWM-III produces accuracies that are superior to the other four term weight measures over the full range of various numbers of features in both models. The TWM-III shows accuracy reduction when the number of features changes from 6000 to 7000 in BOW model, where as in PDW model the accuracy consistently increases when the number of features increases. The TWM-III produces a good accuracy of 84.19% for location prediction in PDW model which is far better than the accuracy (70.90%) in BOW model.

6.4 Discussion

The term frequency (TF) alone may not have enough discriminative power to differentiate the writing styles of the authors in Author Profiling. In TFIDF measure, the IDF measure assigns more weight to the terms which are occurred in less number of documents. It was observed that only term frequency and document frequency are not sufficient in differentiating the profiles of the authors

In our reviews dataset, most of the terms which are considered in vocabulary set are equally distributed in male and female documents. When calculating the tf * rf measure for these terms using equation (2), the values of *a* and *c* are tending to be similar. As a result, these terms are not useful to discriminate the above documents.

In hotel reviews, most of the terms such as room, food, service etc., commonly occur in both male and female documents. These terms are not influence the discrimination of above documents since the values of $|C|$ and cf are tending to be similar and the value of the measure $N/(tp + fn)$ is approximately becomes to 1 in equation (3).

The proposed term weight measure as in equation (4) computes the term weights based on the term frequency in a specific profile group, document importance, document frequency in a interested profile group and document frequency in other profile groups. The measure $tf(t_i, d_k)/ tf(t_i, p_j)$ represent the term importance in a specific profile group of documents. The document importance is estimated based on the information in it. The measure $log(d_{tk})/(0.8*AVGUT_k+0.2*UT_k)$ indicates the document importance. In this measure number of

unique terms is used to specify the importance of the document. If the document contains more number of unique terms then the document contains more information. The measure $a_{ij} / (a_{ij}+b_{ij})$ gives the strength of the term t_i in profile group p_j and $c_{ij} / (c_{ij}+d_{ij})$ gives the importance of the term t_i in other profile groups except p_j.

Table 3. Performance of BOW and PDW model with proposed term weigh measure using NBM and RF algorithms are used for gender, age and location prediction

Profiles	Models/ Classifiers	BOW Model	PDW Model
Gender	NBM	84.20 %	92.37 %
	RF	82.30 %	93.15 %
Age	NBM	75.95 %	83.31 %
	RF	78.30 %	86.97 %
Location	NBM	67.85 %	81.49 %
	RF	70.90 %	83.97 %

The performance of BOW model and PDW model using NBM and RF algorithms for gender, age and location are represented in Table 3. From the results, it shows that the proposed term weight measure consistently performs well compared to other four term weight measures when the number of features was increased. It was observed that the unsupervised term weight methods are not showing a consistent performance, but the proposed supervised method have shown best performance for predicting the profiling characteristics of the authors. It was also observed that as the number of features increases, there is an improvement in the accuracy of the term weight measures in the proposed PDW model. This is because a larger number of features are more likely to cover all profile groups and in turn better able to discriminate between documents in different profile groups.

The proposed term weight measure (TWM-III) obtained good accuracies when Naïve Bayes Multinomial (NBM) algorithm used compared to Random Forest (RF) algorithm for gender prediction, but RF algorithm perform well compared to NBM for age and location prediction.

7. Conclusions and Future Scope

Term weights are playing an important role for constructing document vectors. Different term weight methods, including unsupervised and supervised term weighting approaches, have been extensively investigated in different information retrieval applications. The increased popularity of the Author Profiling, several researchers proposed various solutions to predict the profiles of the authors. On the reviews dataset, all the five term weight measures produce poor results when the number of features is low (1000). TWM-I, TWM-II, TWM-III obtain their best accuracy when the number of features are increased from 1000 to 8000. The proposed term weight measure and proposed model perform well to increase the accuracies of the gender, age and location. In the proposed model, the term to document correlation is achieved by calculating document weight with term weights in that document and document to profile correlation is achieved by generating document vectors with document weights and profile label. The proposed approach performed well than existing state of the art approaches.

It is planned to propose a new supervised term weight measures to increase the accuracy, extend this work on various datasets and prediction of other profiles of the authors. In addition to the above it is planned to modify Random Forest machine learning algorithm to increase the accuracy of the profiles prediction.

References

[1] K.S. Jones, “A Statistical Interpretation of Term Specificity and Its Application in Retrieval,” J. Documentation, vol. 28, no. 1, pp. 11-21, 1972.

[2] M. Koppel, S. Argamon and A. Shimoni, “Automatically categorizing written texts by author gender”, Literary and Linguistic Computing, pages 401-412, 2003.

[3] 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, pp. 199-205, March 2006.

[4] J. Pennebaker “The Secret Life of Pronouns: What Our Words Say About Us. Bloomsbury“, USA 2013.

[5] J.S. Company, L. Wanner. “How to Use Less Features and Reach Better Performance in Author Gender Identification”. The 9th edition of the Language Resources and Evaluation Conference (LREC), pp. 1315-1319, May, 2007.

[6] M. De-Arteaga, S. Jimenez, G. Duenas, S. Mancera and J. Baquero, “Author Profiling Using Corpus Statistics, Lexicons and Stylistic Features”, Proceedings of CLEF 2013 Evaluation Labs, 2013.

[7] W.Y. Lim, J.Goh and V.L.L. Thing, “Content-centric age and gender profiling”, Proceedings of CLEF 2013 Evaluation Labs, 2013.

[8] S. Mechti, M. Jaoua, L.H. Belguith, “Author Profiling Using Style-based Features”, Proceedings of CLEF 2013 Evaluation Labs, 2013.

[9] S. Maharjan, P. Shrestha, and T. Solorio, “A Simple Approach to Author Profiling in MapReduce”, Proceedings of CLEF 2014 Evaluation Labs, 2014.

[10] A. Grivas, A. Krithara, and G. Giannakopoulos, “Author Profiling using stylometric and structural feature groupings”, Proceedings of CLEF 2015 Evaluation Labs, 2015.

[11] A.P. Garibay, A.T.C. Gonzalez, R.A.F. Villaneda, I.H. Farias, D. Buscaldi and I.V.M. Ruiz, “A Random Forest Approach for Authorship Profiling”, Proceedings of CLEF 2015 Evaluation Labs, 2015.

[12] O.M. Sulea and D. Dichiu, B. Romania, “Automatic Profiling of Twitter Users Based on Their Tweets.”, Proceedings of CLEF 2015 Evaluation Labs, 2015.

[13] T. Raghunadha Reddy, B.VishnuVardhan, and P.Vijaypal Reddy, “A Survey on Authorship Profiling Techniques”, International Journal of Applied Engineering Research, Volume 11, Issue 5, pp 3092-3102, march 2016.

[14] D. Estival, T. Gaustad, S.B. Pham, W. Radford and B. Hutchinson, “Author Profiling for English Emails”. 10th Conference of the Pacific Association for Computational Linguistics (PACLING, 2007), pp 263-272, 2007.

[15] P.D. Duc, T.G. Binh, P.S. Bao, “Author Profiling for vietnamese blogs”, Asian Language Processing, 2009 (IALP '09), pp. 190-194. (2009).

[16] P.D. Duc, T.G. Binh, P.S. Bao, “Authorship Attribution and Gender Identification in Greek Blogs”, 8th International Conference on Quantitative Linguistics (QUALICO), pp. 21-32, April 26-29, 2012, (2012).

[17] S. Argamon, M. Koppel, J.W. Pennebaker, and J. Schler (2009), “Automatically profiling the author of an anonymous text”, Communications of the ACM, 52(2), pp. 119-123, Feb 2009.

[18] M. Lan, C.L. Tan, J. Su, and Y. Lu, “Supervised and Traditional Term Weighting Methods for Automatic Text Categorization,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 31, no. 4, pp. 721-735, Apr. 2009.

[19] X. Quan, W. Liu, and B. Qiu, “Term Weighting Schemes for Question Categorization”, IEEE transactions on pattern analysis and machine intelligence, vol. 33, no. 5, pp. 1009-1021, may 2011.