Indian Journal of Science and Technology, *Vol 9(46), DOI: 10.17485/ijst/2016/v9i46/99404, December 2016*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Author Profile Prediction using Pivoted Unique Term Normalization

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

[1]Department of IT, Vardhaman College of Engineering, Shamshabad, Hyderabad - 500018, Telangana, India; trnreddy543@gmail.com
[2]Department of CSE, JNTUH College of Engineering, Karimnagar - 505501, Telangana, India; mailvishnu@jntuh.ac.in
[3]Department of CSE, Matrusri Engineering College, Hyderabad - 500059, Telangana, India; drpvijayapalreddy@gmail.com

Abstract

Objective: Author Profiling is a text classification technique to predict the author profiles of anonymous text. Author Profiles are the demographic characteristics of the authors like age, gender, native language, location, educational background and personality traits. This paper proposes a new model to predict the profiles of the authors such as gender and age by analyzing their writing styles on hotel reviews dataset. **Method:** Most of the existing approaches suffer from high dimensionality of features and capturing the relationship between the features. In this paper, a Profile specific Document Weighted approach is proposed to address the drawbacks of existing approaches. In the proposed model, the pivoted unique term normalization measure is used to calculate the weight of the terms specific to each profile group. Document weight specific to each profile group is calculated as the sum of individual term weights of the specific group. A document vector has constructed using the weight of each group of a profile in the document. **Findings**: An anonymous document profile has identified using the model generated by the machine learning classifier. The performance of the proposed model is evaluated with various classifiers using accuracy as a measure. The proposed approach is experimented on reviews domain to predict the gender and age group of the authors. For gender prediction, the proposed model trained on Naïve Bayes Multinomial classifier results to good accuracy of 91.50%. The logistic classifier results to good accuracy of 81.58% on the proposed model. The results achieved in this paper that outperforms most of the existing approaches. **Applications:** Author Profiling became popular in several information technologies enabled applications such as marketing, forensic analysis, psychology and entertainment. Using reviews dataset on hotels, the business analysts can take strategic decisions to improve the business by identifying the individual profile groups based on the customer reviews.

Keywords: Accuracy, Age Prediction, Author Profiling, Document Weight, Gender Prediction, Pivoted Unique Term Normalization

1. Introduction

In recent times, the Internet has suffered from the exponential growth of publicly available textual data generated by different users, mainly through reviews, blogs and social media. The extraction of valuable information from this huge amount of data has attracted the attention of the researchers from different areas. In this context, Author Profiling is a technique that is concentrated by several researchers to extract key information from the text by analyzing the text itself.

Author Profiling is an important technique in the present information era which has applications in

**Author for correspondence*

marketing, security and forensic analysis. Social websites are an integral part of our lives through which, crimes are cropping up like public embarrassment, fake profiles, defamation, blackmailing, stalking etc. To identify the perpetrator, it is useful by understanding the writing style of perpetrator using Author Profiling. Forensics is a field to analyze the style of writing, signatures, documents, and anonymous letters to identify the terrorist organizations. In the marketing domain, the consumers were provided with a space to review the product. Most of the reviewers were not comfortable in revealing their personal identity. In this case these reviews were analyzed to classify the consumers based on their age, gender, occupation, nativity language, country and personality traits. 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. It helps in revealing the exceptional talent of the students. It also helps in estimating the suitable level of knowledge of each student or a student group in the educational forum.

In general, every human being has his own style of writing and the writing style will not be changed while writing in Twitter tweets, blogs, reviews, social media and also in documents. According to[1], men use number of determiners, prepositions 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.

The content based features are more useful to distinguish the writing styles of male and female authors. The occurrence of words like world cup and cricket increases the chances of text written by male and the occurrence of words such as my husband, pink and boyfriend increases the chances of text written by female and observed that the male authors are tend to use more prepositions in their writings when compared to female 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, favorite heroines/heroes and college life and the users belonging to 33-47 age group post more about post-marriage life and corporate/social activities[2].

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[3], it is informed that females write more about wedding styles and males write more about technology and politics. Further females use more adjectives and adverbs than male authors. They also observed that the content based features alone are more discriminative for gender and age prediction than the rest of the features and they observed that a slight decrease in accuracy when the content based features were added to other features.

In[4], observed that females are more likely to include verbs, negations, pronouns, words related to home, friends, family and various emotional words. Males tend to use more number of articles, prepositions, numbers and longer words. In another observation[5], that the number of prepositions and determiners usage was increased with age, as well as the number of pronouns and negations usage was decreased with age. The older authors write longer posts by using longer words and they concentrated more on usage of commas in their writings and the younger authors use more pronouns, less nouns and articles[6]. [7]conducted a survey on all the practices and issues related to document management and observed that the implementation of practices of document should be project specific rather than organization specific.

The main focus of this paper is to predict the gender and age group of the authors in reviews domain by exploiting the writing styles of the authors. This paper is organized as follows: section 2 explains the existing approaches for Author Profiling. The proposed work is described in section 3. The achieved results are analyzed in section 4. Section 5 concludes this work.

The most straightforward approach for representing the document text is by the vector of word frequencies. This phenomenon is similar to the conventional Bag Of Words (BOW) representation and then the researchers were concentrated on the topic based classification and observed that the best accuracies for age and gender prediction were achieved when context information of the blog was used[8]. Bag Of Words (BOW) is one of the common approach that builds the feature vectors of documents by taking every term in the vocabulary as an attribute. In the text classification on topics, the common words such as pronouns, articles, prepositions were generally removed from the feature set because of the fact that they do not possess any semantic relationship and are termed as function words. In general, the text is defined as a set of words and every word possess some frequency without focusing on the contextual information. Style based text classification is observed to possess a significant

difference where in the best features were found to be the most common words that are to discriminate between the authors. In [9], observed that the current benchmark features that rely on term frequency are not more suitable for document clustering and they considered various concept, semantic and context relations during document clustering.

The size of the feature terms places a predominant role in the document representation. Few researchers[10,11] used 3000 frequent terms as the size of feature set and incremented up to 50000 frequent terms. The researchers also deduced that not only the features set size but the classification algorithm also has a significant role when the dimensionality of the problem increases which over fits the training data. When the size of the feature set is more, various researchers used different feature selection techniques to reduce the size of the feature set. in [11]followed some ideas from Concise Semantic Analysis (CSA) to use a low dimensional representation with high level of representativeness. In[12], Applied Principal Component Analysis (PCA) as a method which transform the high dimensional data into a lower dimensional linear space for simple representation of the data.

In[13] used corpus of blog posts of 19320 blog authors. They achieved good accuracies for prediction of age and gender of the blog posts by using both content based features of n-grams and stylistic features. In their observation the style based features like determiners, prepositions and pronouns were most useful to discriminate the male and female authors. In[14], experimented with n-gram based approach to detect the plagiarism in Hindi documents. They divided the text into n-grams and cosine similarity measure was used to find the matching percentage of given document and repository of documents. In in[15] enhanced the existing algorithm of Enhance Concept based Similarity Measure for Text Processing (ECSMTP) to find the similarity between multiple documents by comparing line by line of the documents.

In[16] proposed an approach, where in all the training data documents are indexed by information retrieval engine and then they treated every test document as a query. They used simple information retrieval features to predict gender and age from social media texts. In another experiment[17], they increased the number of features to 64 including information retrieval features. In[18], it was applied same information retrieval features to predict various profiling characteristics of the authors and observed that these features are suitable for predicting personality traits of the authors.

In[19], collected 9836 emails of 1033 authors. In their work 689 features of character level, lexical and structural features were used. Various machine learning algorithms including J48, RandomForest, IBK, JRip, SMO, libSVM, Bagging, AdaBosstM1 were applied on the corpus. Among all these classifiers SMO results best accuracy for gender and age prediction. In[20], observed that the classification algorithms along with clustering algorithms are increased the accuracy for document summarization based on the word classification.

In[21], experimented on the corpus of 1672 texts of New York Times opinion blogs. They tried different combinations of word based, character based, sentence based, dictionary based and syntactic features for gender prediction. The better results achieved when all the features were considered. They observed that the accuracy is reduced when the Bag of Words approach is applied on this corpus with 3000 words having most tf-idf values.

In[22] collected 3524 pages of 73 Vietnamese bloggers. 298 features including word based and character based features were considered in their work and found that word based features contributed more to predict gender than character based features. In their experiment, 10 machine learning algorithms from weka toolkit were used namely ZeroR, Decision Tree J48, RandomForest, Bagging, IBK, SMO, NiveBayes, BayesNetwork, Multilayer Perception, and Random Tree. Among these classifiers, IBK classifier results a good accuracy for gender and age prediction when all features were used together. In[23] used 1000 blog posts of 20 bloggers from Greek language and considered standard stylometric features and 300 most frequent word n-grams and character n-grams. Support Vector Machine is trained on this corpus and generated good accuracy for gender prediction and realized that longer sequences of word n-grams and character n-grams increase the prediction accuracy of a gender.

Most of the researchers faced problems in their approaches to Author Profiling like high dimensionality, fails to capture the relationship between features, sparsity in document representation and over fitting of a classifier. In this paper a new approach is proposed to solve the problems of existing approaches in Author Profiling.

2. Proposed Approach

In this approach, initially the term weights were calculated specific to profiles by using pivoted document length normalization measure. Second, the document weights were calculated specific to profiles by aggregating the weights of terms in that document and finally these weights were used to create a document vectors for designing a classification model. The procedure of the proposed approach was showed in Figure 1.

In this model $\{T_1,T_2,\ldots..T_n\}$ denotes the collection of vocabulary terms, $\{D_1,D_2,\ldots,D_m\}$ is a collection of documents in the corpus. TWM is a term weight in male corpus, TWF is a term weight in female corpus, TW18-24 is a term weight in a 18-24 age group corpus, TW25-34 is a term weight in a 25-34 age group corpus, TW35-49 is a term weight in a 35-49 age group corpus, TW50-64 is a term weight in a 50-64 age group corpus and TW65_AND_ABOVE is a term weight in a 65_AND_ABOVE age group corpus. DWM represents the weight of a document in male corpus, DWF represents the weight of a document in a female corpus, DW18-24 represents the weight of a document in the age group of 18-14 corpus, DW25-34 represents the weight of a document in the age group of 25-34 corpus, DW35-49 represents the weight of a document in the age group of 35-49 corpus, DW50-64 represents the weight of a document in the age group of 50-64 corpus, DW65_AND_ABOVE represents the weight of a document in the age group of 65_AND_ABOVE corpus.

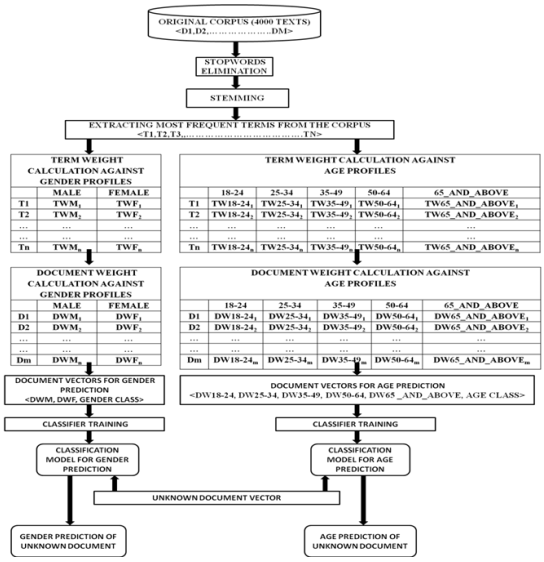

Figure 1. The proposed model for gender and age prediction

The procedure of the proposed approach is explained in the following steps:

Step 1: Collect the Reviews corpus.
Step 2: Preprocess collected corpus for stop words removal and stemming.
Step 3: Extraction of the most frequent terms from preprocessed corpus.
Step 4: The term weights are calculated on author profiles.
Step 5: The Document weights are calculated on author profiles by aggregating weights of terms in that document.
Step 6: Generating document vectors by using weights of the documents.
Step 7: Train the model with document vectors.
Step 8: The test documents are passed to the generated classification model to find the demographic characteristics of the author unknown documents.

The subsection 3.1 describes the relationship between the terms and the profiles. The relationship between the documents and the profiles were described in subsection 3.2.

2.1 Term Representation Specific to Profiles

Term weighting is an important concept in the modern information analysis. Different terms have different importance in a text. The term weight measure is used to find the importance of a term in a text. In general Author Profiling techniques easily analyze and predict the demographics of authors when the document contains large amount of text. For small documents it is difficult to predict the features of the authors. In this paper pivoted Document length normalization technique is used to remove the difficulty of analyzing small sized texts. The document length normalization technique maintains the term weights for a document in accordance with its weight.

Let $P = \{p_1,p_2,\ldots.p_q\}$ is the set of profiles, $\{D_1,D_2,\ldots,D_m\}$ is a collection of documents in the corpus, $V = \{t_1,t_2,\ldots,t_n\}$ is a collection of vocabulary terms for analysis. Each term $t_i \in V$ is represented as a vector t_{ij}, i.e., $t_{ij} = \{t_{i1},t_{i2},\ldots,t_{iq}\}$, where the dimension t_{ij} represents the term t_i weight on the profile p_j. Equation (1) is used to calculate the term weight in a specific profile.

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

where, $W(t_i, p_j)$ is the weight of i^{th} term in j^{th} profile. TF_i (Term Frequency) is the number of times the term t_i

is occurred in a document k, $AVGTF_i$ is a ratio of the term frequency t_i to the total number of terms in k^{th} document. As the experiment performed in pivoted unique normalization[24], the constant value 0.2 for slope is effective across prediction of demographic features of the authors. UT_k is a number of unique terms in k^{th} document, and $AVGUT_k$ is a ratio of number of unique terms to total number of terms in k^{th} document.

The following equations (2.1) and (2.2) are used to normalize the term weight values.

$$\bar{t}_{ij} = \frac{w(t_i, p_j)}{\sum_{i=1}^{n} w(t_i, p_j)} \quad (2.1) \qquad t_{ij} = \frac{\bar{t}_{ij}}{\sum_{j=1}^{q} w(t_i, p_j)} \quad (2.2)$$

where, $\bar{t}_{ij}$ is the ratio of weight of term t_i in profile p_j to the weights of all the vocabulary terms in profile p_j . t_{ij} is the ratio of $\bar{t}_{ij}$ to the weight of the term t_i in all the considered profiles.

2.2 Document Representation Specific to Profiles

In this work, each document is represented by aggregating the weights of terms specific to that document. Equation (3) is used to calculate the weight of document on each profile. In equation (3), Term Frequency Inverse Document Frequency (TFIDF) measure is used to calculate the weight of the term in a particular document. TFIDF (equation (4)) measure gives the weight to a term based on the number of documents contains the same term with respect to a total corpus of documents.

$$W_{dkj} = \sum_{t \in d_k, d_k \in p_j} TFIDF(t_i, d_k) \cdot W_{tij} \quad (3)$$

$$TFIDF(t_i, d_k) = tf(t_i, d_k) * \log\left(\frac{|D|}{|1 + DF_{ti}|}\right) \quad (4)$$

where, W_{dkj} is the weight of document d_k in the profile p_j , $|D|$ is the total number of documents in the respective profile, DF_{ti} is the number of documents contains the term t_i in the profile p_j .

The collections of training documents were finally represented using equation (5)

$$Z = \bigcup_{d_k \in p_j} (z_k, c_j) \quad (5)$$

Here, $z_k = \{W_{dk1}, W_{dk2}, \ldots\ldots, W_{dkq}\}$ and c_j is a class label of profile p_j .

The vector Z contains weights of a document specific to each profile with document profile label.

3. Empirical Evaluations

3.1 Corpus Characteristics

The corpus was collected from TripAdvisor.com, which contains 50000 reviews about different hotels. The corpus was constructed carefully to ensure its quality with regard to text cleanliness and annotation accuracy which resulted with 4000 reviews only. In order to make this dataset applicable to Author Profiling and to ensure its quality, the following steps are adopted. First, reviews containing less than 5 lines of text were excluded from our dataset. Second, the reviews which are written in English language were only considered. Third, the reviews were considered written by the authors whose gender was given in their user profile. After collecting the reviews two preprocessing steps were performed on the corpus such as stop words removal and stemming. In this experiment it was adopted the stop word list as in[25] and the stemming is performed by using porter stemming algorithm[26]. Two gender classes were considered, namely male and female and five classes considered for age such as 18-24, 25-34, 35-49, 50-64 and 65_AND_ABOVE. The corpus is balanced in terms of gender dimension but unbalanced in terms of age dimension, where the amount of users from 18-24 and 65_AND_ABOVE groups were significantly smaller than the amount of users from the rest of the age groups.

Table 1 shows the characteristics of the corpus used for gender and age prediction.

3.2 Evaluation Measures

Most of the researchers used various measures (Precision, Recall, F1-score and Accuracy) for describing their system efficiency. In this work accuracy measure is used to measure the performance. Accuracy is considered as the scoring metric to evaluate the effectiveness of the system.

Table 1. Corpus characteristics

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_above	600	300	300
TOTAL		4000	2000	2000

Accuracy in this context is the ratio of number of correct age or gender predictions in the documents to the total number of documents in the corpus.

$$Accuracy = \frac{Number\ of\ \text{documents correctly predicted their}\ gender\ /\ age}{Total\ number\ of\ documents}$$

3.3 Result Analysis

In this experiment 10-fold cross validation is used to evaluate each feature subset. In 10-fold cross validation, the original corpus is randomly partitioned into 10 subsamples. Of the 10 subsamples, 9 subsamples are used as training data and the remaining one subsample is selected as the validation data for testing the model. The cross validation process is repeated until every subsample is used exactly once as the validation data. In this work, various classifiers such as Naive Bayes Multinomial (Probabilistic), Simple Logistic and Logistic (functional), IBK (lazy), Bagging (Ensemble/meta) and RandomForest (Decision Tree) were used from WEKA tool. The output of a feature extraction was represented as an ARFF file where in the documents were represented as a multidimensional vector. Each feature value is a dimension in a multidimensional vector. The subsection 4.3.1 explains the results for gender prediction and 4.3.2 explains results for age prediction.

3.3.1 Gender Prediction

Various researchers of Author Profiling used different types of features in their experiments namely lexical features, character based features, syntactic features, structural features, semantic features, readability features and information retrieval features. Several techniques to Author Profiling used this combination of features to predict the demographic characteristics of the authors. Most of the researchers are concentrated on the representation of the document with these combinations of features and only few researchers used different representation for a document. In this approach a new document representation is proposed in which the number of features was used to represent a document depends on the range of values of the profile.

In this approach, initially the most frequent terms were identified in the total corpus, then these terms weights are calculated on author profiles. These term weights are aggregated to calculate the weight of a document against profiles. The weight of a document depends on the weights of the terms in that document. Many term weight measures are proposed by several researchers, in this work a pivoted document length normalization weight measure is used to find the weight of the considered terms. This weight measure was used the term frequency and unique terms in a document to find the term weight. The number of unique terms plays a major role in differentiating the writing style of the authors. In general the females write large size of reviews on products than males and the number of unique terms decreased with by increasing the size of the document. The proposed approach achieved good accuracies in gender prediction because of using this effective term weight measure.

The classifiers play a role in accuracy prediction. The proposed approach achieved a good accuracy of 84.55% than the accuracy of[21] approach when bagging classifier was used. In[22] achieved an accuracy of 83.34%, which is less than the accuracy of proposed approach when IBK classifier was used. In this approach Naive Bayes Multinomial classifier gave a best accuracy among other classifier. The Naive Bayes Multinomial classifier is a more accurate classifier for corpuses that have a huge number of documents and have a large variance in lengths of documents. Naive Bayes Multinomial classifier works very fast and is a specialized version of Naive Bayes Classifier. Table

2 shows the comparison of proposed model with existing approaches in Author Profiling. The proposed approach used only two features for representing a document where as other approaches used more number of features to represent a document and best accuracy achieved for gender prediction than most of the approaches existing in Author Profiling.

For gender prediction, only two features were used to represent a document. <DWM, DWF, GENDER CLASS> is the representation of a document vector for gender prediction. Here, DWM is the weight of a document on male profile, DWF is the weight of same document on female profile, GENDER CLASS is the label of a document that is either male or female. The training set contains a set of vectors for each training document. These vectors are used to train the classifiers to identify the profile of a unknown document. The machine learning algorithms were applied in this work which include Naive Bayes multinomial, functions classifiers (Logistic, Simple Logistic), lazy classifier (IBK), ensemble/meta classifier (bagging), and decision tree classifier (RandomForest). The gender prediction was evaluated as a classification problem and accuracy measure was used to report the results. The results achieved for gender prediction in the proposed approach were presented in Table 3.

In Table 3, it is observed that as the number of terms increased for profile specific document representation from 1000 to 8000 with an interval of 1000 words, the growth rate in accuracy was decreased. The Naïve Bayes Multinomial classifier achieved a highest accuracy of 91.5% accuracy for gender prediction by using 8000 most frequent words as terms. It is also witnessed that in all the classifiers the accuracies were increased when the number of features were increased. The proposed approach was given better results than most of the existing approaches on Author Profiling. In all the iterations the Naive Bayes Multinomial classifier proved to perform well than all other classifiers.

3.3.2 Age Prediction

The comparisons of various approaches with proposed model for age prediction in Author Profiling were represented in Table 4. For age prediction, only five features were used to represent a document vector where as other approaches used more number of features to represent a document. <DW18-24, DW25-34, DW35-49, DW50-64, DW65_AND_ABOVE, AGE CLASS> is the representation of a document for age prediction. Here, DW18-24 is the weight of a document on 18-24 profile, DW25-34 is the weight of a document on 25-34 profile, DW35-49 is the weight of a document on 35-49 profile, DW50-64 is the weight of a document on 50-64 profile, and DW65_AND_ABOVE is the weight of a document on 65_AND_ABOVE profile. In [22] achieved an accuracy of 77.27% for gender

Table 2. Comparison of proposed model with existing approaches for gender prediction

APPROACH	CORPUS	NUMBER OF FEATURES	AUTHOR PROFILE	CLASSIFIER	ACCURACY
19.	Emails	689	Gender	SMO	69.26
21	Newyork Times Opinion Blog	83	Gender	Bagging	82.83
22	Vietnamese blogs	298	Gender	IBK	83.34
23	Greek blogs	1356	Gender	SMO (Sequential Minimal Optimization)	82.6
1	British National Corpus	1081	Gender	Exponential Gradient Algorithm	80
6	Blogs	1000	Gender	Bayesian Multinomial Regression	76.1
2	Blogs	1502	Gender	Multi Class Real Winnow Algorithm	80.1
Proposed	Reviews	2	Gender	Naivebayes Multinomial	**91.5**

Table 3. The accuracies of gender prediction for various machine learning classifiers

CLASSIFIER/ NUMBER OF TERMS	NAIVEBAYES MULTINOMIAL	SIMPLE LOGISTIC	LOGISTIC	IBK	BAGGING	RANDOM FOREST
1000 WORDS	79.25 %	76.00 %	79.15 %	69.35 %	72.80 %	72.20 %
2000 WORDS	82.15 %	79.05 %	81.95 %	73.60 %	74.20 %	75.35 %
3000 WORDS	84.60 %	81.75 %	84.25 %	75.45 %	76.70 %	76.60 %
4000 WORDS	86.35 %	83.95 %	86.30 %	79.75 %	78.55 %	79.60 %
5000 WORDS	87.80 %	85.00 %	87.55 %	80.35 %	80.35 %	81.90 %
6000 WORDS	89.75 %	87.60 %	89.05 %	82.85 %	81.90 %	83.65 %
7000 WORDS	90.70 %	88.70 %	89.90 %	85.55 %	83.60 %	85.05 %
8000 WORDS	**91.50 %**	90.00 %	90.85 %	85.80 %	84.55 %	86.50 %

prediction, which is more than the proposed approach when IBK classifier was used. Overall the proposed approach achieved good accuracy for age prediction than that of the existing approaches in Author Profiling.

Table 5 shows the accuracies of age prediction. The logistic classifier achieved a highest accuracy of 81.58 % for age prediction among other classifiers by using 8000 most frequent words as terms. Logistic classifier is popular and powerful classifier. This classifier used logit transform to predict probabilities directly. Logistic classifier fits for a full multinomial logistic regression model subject to the condition that all attributes uses a ridge estimator.

4. Conclusions and Future Scope

In this paper a new document representation was proposed for Author Profiling in reviews domain. The proposed approach captures the term to profiles and the document to profiles relationship information in non-sparse and low dimensional vector space. This model is language independent. The proposed approach is adoptable for any type of feature sets and any type of classification model. In this work an overall accuracy of 91.50% for gender prediction and 81.58 % for age prediction was obtained. In general the approaches for Author Profiling achieve higher accuracy for gender prediction than age prediction because the range of values for gender dimension was less when compared to the range of values of age dimension. This is the accuracy which is proved to be above the range of the accuracies achieved by the various other approaches in this area. However, it is remarkable that this accuracy is achieved with minimal number of features than in most of the state-of-the-art approaches.

Table 4. Comparison of proposed model with existing approaches for age prediction

APPROACH	CORPUS	NUMBER OF FEATURES	AUTHOR PROFILE	CLASSIFIER	ACCURACY
19	Emails	689	Age	SMO	56.46 %
22	Vietnamese Blogs	298	Age	IBK	77.27 %
6	Blogs	1000	Age	BMR (Bayesian Multinomial Regression)	77.7 %
2	Blogs	1502	Age	Multi Class Real Winnow algorithm	76.2 %
Proposed	Reviews	5	Age	Logistic classifier	81.58 %

Table 5. The accuracies of age prediction for various machine learning classifiers

CLASSIFIER/ NUMBER OF TERMS	NAIVEBAYES MULTINOMIAL	SIMPLE LOGISTIC	LOGISTIC	IBK	BAGGING	RANDOM FOREST
1000 WORDS	47.07 %	34.23 %	54.15 %	31.35 %	33.80 %	34.81 %
2000 WORDS	54.34 %	39.56 %	62.98 %	36.56 %	38.00 %	40.51 %
3000 WORDS	59.24 %	44.35 %	67.33 %	43.67 %	41.99 %	45.11 %
4000 WORDS	66.12 %	48.21 %	72.33 %	45.11 %	44.53 %	50.08 %
5000 WORDS	69.87 %	52.38 %	74.99 %	49.25 %	47.13 %	52.77 %
6000 WORDS	73.64 %	57.03 %	78.06 %	53.54 %	51.12 %	57.43 %
7000 WORDS	76.25 %	59.85 %	79.89 %	55.99 %	52.47 %	61.30 %
8000 WORDS	78.18 %	62.06 %	81.58 %	56.94 %	54.40 %	61.66 %

In our future work, it is planned to expand this method to identify the demographic features of authors like native language, location, educational background and personality traits. It is also planned to extract new features to increase the accuracies of gender and age prediction.

5. References

1. Koppel M, Argamon S, Shimoni A. Automatically categorizing written texts by author gender. Literary and Linguistic Computing; 2003. p. 401–12.
2. Schler J, Koppel M, Argamon S, Pennebaker J. Effects of age and gender on blogging. Proceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs; 2006 Mar.
3. Pennebaker J. The secret life of pronouns: What our words say about us. Bloomsbury, USA; 2013.
4. Newman ML, Groom CJ, Handelman LD, Pennebaker J. Gender differences in language use. An Analysis of 14,000 Text Samples Discourse Processes; 2008. p. 211–36.
5. Pennebaker JW, Francis ME, Booth RJ. Linguistic inquiry and word count: LIWC 2001. Mahway: Lawrence Erlbaum Associates; 2001.
6. Argamon S, Koppel M, Pennebaker JW, Schler J. Mining the blogosphere: Age, gender and the varieties of self-expression. First Monday. 2007; 12(9).
7. Satish CJ, Anand M. Software documentation management issues and practices: A Survey. Indian Journal of Science and Technology. 2016 May; 9(20):1–7.
8. Santosh K, Romil B, Shekhar M, Vasudeva V. Author profiling: Predicting age and gender from blogs. Proceedings of CLEF 2013 Evaluation Labs; 2013.
9. Rao AS, Ramakrishna S, Babu PC. MODC: Multi-Objective Distance based Optimal Document Clustering by GA. Indian Journal of Science and Technology. 2016 Jul; 9(28):1–8.
10. López-Monroy AP, Montes-y-Gómez M, Hugo JE, Luis VP. Using intra-profile information for author profiling. Proceedings of CLEF 2014 Evaluation Lab; 2014.
11. López-Monroy AP, Montes-y-Gómez M, Hugo JE, Luis VP, Esaú VT. INAOE's participation at PAN'13: Author profiling task. Proceedings of CLEF 2013 Evaluation Labs; 2013.
12. Wee-Yong L, Jonathan G, Vrizlynn LLT. Content-centric age and gender profiling. Proceedings of CLEF 2013 Evaluation Labs; 2013.
13. Argamon S, Koppel M, Pennebaker JW, Schler J. Automatically profiling the author of an anonymous text. Communications of the ACM. 2009; 52(2):119.
14. Urvashi G, Vishal G. Maulik: A plagiarism detection tool for Hindi documents. Indian Journal of Science and Technology. 2016 Mar; 9(12):1–11.
15. Kalpana S, Vigneshwari S. Selecting multiview point similarity from different methods of similarity measure to perform document comparison. Indian Journal of Science and Technology. 2016 Mar; 9(10):1–6.
16. Edson RDW, Viviane PM, José PMO. Exploring information retrieval features for author profiling. Proceedings of CLEF 2014 Evaluation Labs; 2014.
17. Edson RDW, Viviane PM, José PMO. Using simple content features for the author profiling task. Proceedings of CLEF 2013 Evaluation Labs; 2013.
18. Edson RDW. Information retrieval features for personality traits. Proceedings of CLEF 2015 Evaluation Labs; 2015.
19. Estival D, Gaustad T, Pham SB, Radford W, Hutchinson B. Author profiling for english email. 10th Conference of the Pacific Association for Computational Linguistics (PACLING); 2007.
20. Dharinya S. Analysis of document summarization and word classification in a smart environment. Indian Journal of Science and Technology. 2016 May; 9(19):1–7.
21. Juan SC, Leo W. How to use less features and reach better performance in author gender identification? The 9th

edition of the Language Resources and Evaluation Conference (LREC); 2007 May. p. 26–31.
22. Dang DP, Giang BT, Bao PS. Author profiling for Vietnamese blogs. Asian Language Processing (IALP); 2009. p. 190–4.
23. Dang DP, Giang BT, Bao PS. Authorship attribution and gender identification in Greek blogs. 8th International Conference on Quantitative Linguistics (QUALICO); 2012 Apr. p. 26–9.
24. Amit S, Chris B, Mandar M. Pivoted document length normalization. In SIGIR '96: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1996, New York, USA; 1996. p. 21–9.
25. Stopwords list [Internet]. [cited 2000]. Available from: http://members.unine.ch/jacques.savoy/clef/index.html.
26. Porter MF. Developing the English Stemmer; 2002.