International Journal of Engineering

Journal Homepage: www.ije.ir

A Document Weighted Approach for Gender and Age Prediction Based on Term Weight Measure

T. Raghunadha Reddy*[a], B. Vishnu Vardhan[b], P. Vijayapal Reddy[c]

[a] Department of Information Technology,Vardhaman College of Engineering, Hyderabad, Telangana, India
[b] Department of Computer Science and Engineering, JNTUH College of Engineering, Jagtiyal, Karimnagar, Telangana, India
[c] Department of Computer Science and Engineering, Matrusri Engineering college, Hyderabad, Telangana, India

PAPER INFO

Paper history:
Received 02 September 2016
Received in revised form 11 February 2017
Accepted 10 March 2017

Keywords:
Author Profiling
Gender Prediction
Age Prediction
Term Weight
Document Weight
Classifiers

ABSTRACT

Author profiling is a text classification technique, which is used to predict the profiles of unknown text by analyzing their writing styles. Author profiles are the characteristics of the authors like gender, age, nativity language, country and educational background. The existing approaches for Author Profiling suffered from problems like high dimensionality of features and fail to capture the relationship between the features. In this work, a new document weighted approach is proposed in order to address the problems in existing approaches. In this approach, the term weight measure is used to assign suitable weight to the terms and these term weights are aggregated to compute the document weight. The classification model is generated with these document weights for predicting profiles of the text. The proposed approach and existing approaches are experimented on reviews domain with different classifiers. The accuracies of the proposed approach for gender and age prediction are promising than existing approaches.

doi*: 10.5829/idosi.ije.2017.30.05b.03*

NOMENCLATURE

NBM	Naive Bayes Multinomial
SL	Simple Logistic
LOG	Logistic
BAG	Bagging
RF	Random Forest

1. INTRODUCTION

In present days, the Internet has suffered from the incremental growth of publicly available textual data generated by different users, mainly through blogs, social media, twitter tweets and reviews. Most of the users are posting anonymous text in the internet. The extraction of valuable information from this huge amount of anonymous text has attracted the attention of the researchers from different areas. Author profiling is one area that is concentrated by several researchers to extract key information from the text by analyzing the writing styles of the text.

Author Profiling is an important technique in the present information era which has applications in security, forensic analysis, marketing and educational domain [1]. In the marketing domain, the consumers are provided with a space to review the product. Most of the reviewers were not comfortable in revealing their personal identity. These reviews were analyzed to classify the consumers based on their age, gender, occupation, nativity language and country. Based on the

*Corresponding Author's Email: *raghu.sas@gmail.com* (T. Raghunadha Reddy)

Please cite this article as: T. Raghunadha Reddy, B. Vishnu Vardhan, P. Vijayapal Reddy, A Document Weighted Approach for Gender and Age Prediction Based on Term Weight Measure, International Journal of Engineering (IJE), TRANSACTIONS B: Applications Vol. 30, No. 5, (May 2017) 643-651

classification results, companies try to adopt new business strategies to serve the customers.

Generally, every human being has his own style of writing and maintains the same style while writing in Twitter tweets, blogs, reviews, social media and also in documents. According to Koppel et al. [1], the number of determiners and quantifiers usage is more in men writings and the number of pronouns usage more in female writings. Similarly, topics related to sports, politics and technology are more discussed by male authors whereas the topics like beauty, kitty parties and shopping are talk about female authors [1]. Prior works [1, 2] found that more number of prepositions are used by male authors in their articles and blog posts when compared to female authors.

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. 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].

In general, 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], the females write more about wedding styles and males write more about technology and politics. Further females use more adjectives and adverbs than male authors. 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 [4].

Pennebaker et al. [5] observed that the number of prepositions and determiners usage was increased with age, as well as the number of pronouns and negations usage were 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].

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 in five sections. Section 2 explains the existing work in Author profiling. The proposed work is explained in section 3. The corpus characteristics and the comparison of obtained results with existing approaches are described in section 4. Section 5 concludes this work and suggested the future scope in Author profiling.

2. EXISTING WORK

Most of the existing approaches representing the document text is by the vector of word frequencies. This phenomenon is similar to the conventional Bag Of Words (BOW) representation. The BOW approach builds the document vectors by taking every term in the vocabulary list as an attribute. Later, the researchers were concentrated on the topic based classification [7]. In the topic based classification, the common words such as determiners, pronouns, articles and prepositions were generally removed from the feature set because of the fact that they do not helpful to differentiate the writing styles of the authors and are termed as function words.

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 stylistic features. The best accuracies were obtained for age and gender prediction when the content and style based features are used along with context information of the blog [7].

Argamon et al. [8] considered style based features such as function words, part-of-speech and content based features such as 1000 words. They experimented on a dataset of 19320 postings of English blog authors. The classification model is generated by using Bayesian Multinomial Regression and it was observed that the combination of both stylistic and content based features achieved best accuracies for gender and age prediction. The number of features in a vocabulary list places a predominant role in the document representation. Sapkota et al. [9] used the top 5000 words for the prediction of the gender but the results were poor with respect to the English language, but achieved good results on Spanish language.

The researchers also deduced that not only the features set size, the classification algorithm also play a significant role when the dimensionality of the features increased. Hamidi and Daraee [10] reviewed in their study about various applications of classifcation algorithms. When the size of the feature set is more, various researchers used different feature selection techniques to reduce the size of the feature set. Lim et al. [11] applied Principal Component Analysis (PCA) to transform the high dimensional data into a lower dimensional linear space for simple representation of the document. Darvishi and Hassanpour [12] concentrated on several dimensionality reduction techniques like feature extraction techniques and the significance of similarity measures in detecting the relationships between data.

Mechti et al. [13] used TFIDF measure to compute the feature value in representation of a document. They identified a ranked list of words to find the stylistic similarity between male and female. Maharjan et al.

[14] recognized n-grams of words as features and TFIDF measure as the weighting measure. TFIDF scores of the word n-grams were used to filter the n-grams that were not important. Grivas et al. [15] experimented with TFIDF scores of word n-grams to generate feature vectors.

Palomino-Garibay et al. [16] tested 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 et al. [17] used 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.

Weren et al. [18] proposed an approach in which all the training documents are indexed by information retrieval engine and the test document was treated as a query. The simple information retrieval features such as Okapi and BM25 measures were used to predict the gender and age group from social media texts. In another experiment [19] they increased the number of features to 64 including information retrieval features. It was observed that the information retrieval features are suitable for predicting personality traits of the authors compared to gender and age [20].

Estival et al. [21] extracted 689 features such as lexical, character level and structural features from a corpus of 9836 emails of 1033 authors to represent a document vectors. Several machine learning classifiers such as IBK, JRip, SMO, J48, libSVM, RandomForest, Bagging and AdaBoost are used to generate a classification model. Among these classifiers SMO classifier by using all combination of features obtained a good accuracy for age and gender prediction.

Soler and Wanner et al. [22] experimented on the corpus of 1672 texts of New York Times opinion blogs. They extracted different combinations of word based, character based, sentence based, dictionary based and syntactic features for gender prediction. It was observed that the better results achieved when all the features were considered and also observed that the accuracy is reduced when the BOW approach is applied on this corpus with 3000 words having most tf-idf values.

Pham et al. [23] experimented their work on 3524 pages of 73 Vietnamese bloggers. They extracted 298 features such as Lexicon, Character based, Content Specific, Document based, Paragraph based, Word-based, Structural, Line-based, POS-based, Function words to represent the documen vector and applied them on various machine learning algorithms namely Neuron Network (Multilayer Perceptron), IBk (IB1), ZeroR, Bagging, Decssion Tree J4.8, SMO, NaiveBayes, BayesNetwork, Random Forest and RandomTree to generate a classification model. It was observed that the word based features contributed more than character based features to predict gender and IBK classifier results a good accuracy for gender and age prediction when all combination of features were used. In another work, Dang et al. [24] extracted 1000 blog posts of 20 bloggers from Greek language. They considered standard stylometric features and 300 most frequent word n-grams and character n-grams. It was observed that the Support Vector Machine generated good accuracy for gender prediction and also realized that longer sequences of word n-grams and character n-grams increase the prediction accuracy of a gender.

3. PROPOSED APPROACH

Most of the work in author profiling concentrated on the extraction of features, representation of features as document vectors and the machine learning algorithms used for generating classification model. The existing approaches used more number of features to represent a document vector thereby high dimensionality problem is occurred and the features independently participated in the generatiion of classification model thereby no usage of the relationship between features. In this approach, a new document representation is proposed in order to deal with the drawbacks of existing approaches.

In the proposed approach, first preprocessing techniques such as stop word removal and stemming are applied on the collected reviews corpus. From the updated corpus extract most frequent terms that are occurred at least two times in the corpus. Compute the term weights specific to each profile group by using term weight measures. Document weight specific to each profile group is calculated by aggregating the weights of the terms specific to that document using document weight measure. These document weights are used to represent the document vectors. Various classification algorithms are used to generate classification model and this model is used to predict the characteristics of anonymous text.

Figure 1 shows the model for proposed document weighted approach. In this model (D_1,D_2,….,D_m) is a collection of documents in the corpus, (T_1,T_2,….,T_n) denotes the collection of vocabulary terms. TWM, TWF are term weights in a male and female profile group, respectively. DWM, DWF represents the document weight in male and female profile group, respectively. TW18-24, TW25-34, TW35-49, TW50-64 and TW65_AND_ABOVE are the term weights in 18-24, 25-34, 35-49, 50-64 and 65_AND_ABOVE age group corpus respectively. DW18-24, DW25-34, DW35-49, DW50-64, DW65_AND_ABOVE represents the document weights in the age group of 18-24, 25-34, 35-49, 50-64, 65_AND_ABOVE corpus respectively.

The subsection 3.1 describes the term weight computation specific to the profiles. The calculation of document weights specific to the profiles are described in subsection 3.2.

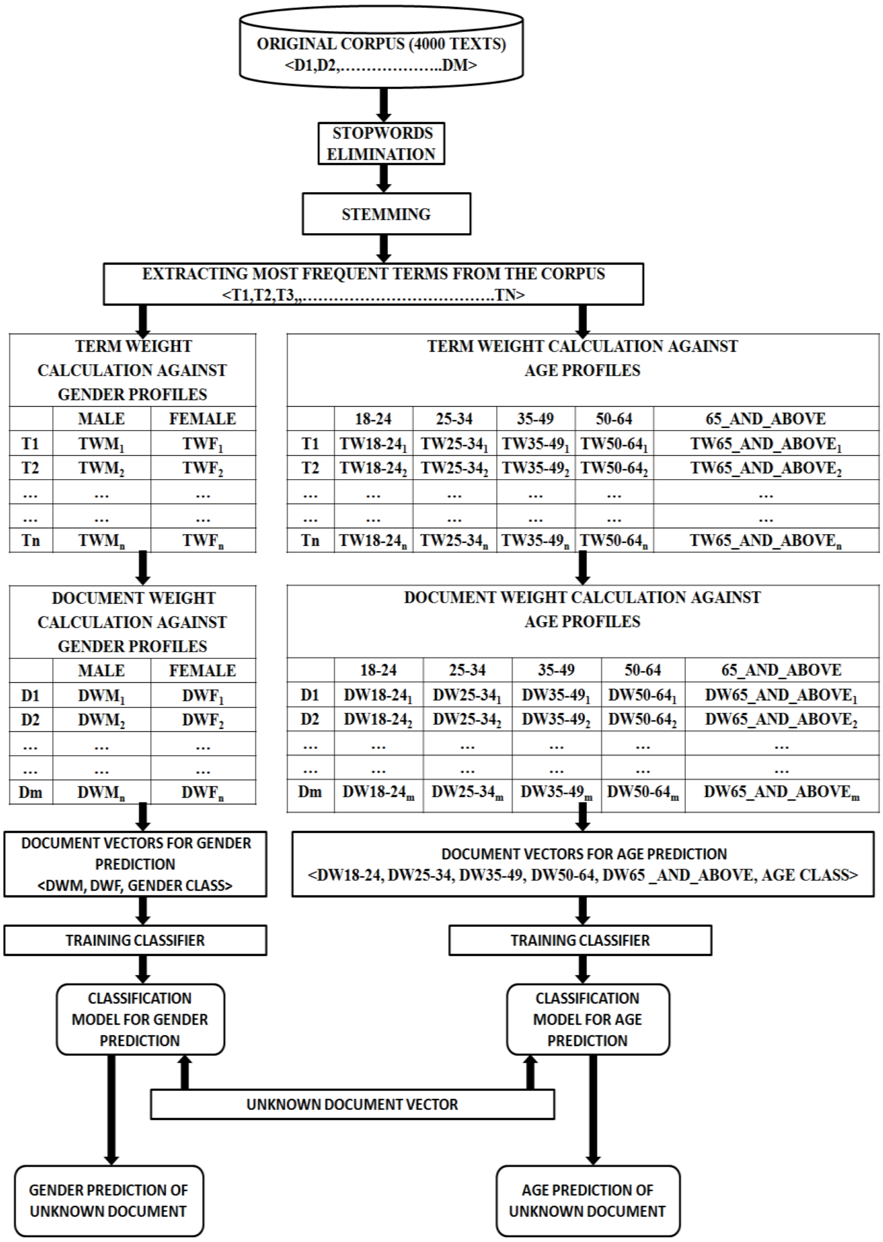
ORIGINAL CORPUS (4000 TEXTS)
<D1,D2,……………….DM>
STOPWORDS ELIMINATION
STEMMING
EXTRACTING MOST FREQUENT TERMS FROM THE CORPUS
<T1,T2,T3,,……………………………..TN>
TERM WEIGHT CALCULATION AGAINST GENDER PROFILES
MALE
FEMALE
T1 TWM1 TWF1
T2 TWM2 TWF2
Tn TWMn TWFn
TERM WEIGHT CALCULATION AGAINST AGE PROFILES
18-24
25-34
35-49
50-64
65_AND_ABOVE
T1 TW18-241 TW25-341 TW35-491 TW50-641 TW65_AND_ABOVE1
T2 TW18-242 TW25-342 TW35-492 TW50-642 TW65_AND_ABOVE2
Tn TW18-24n TW25-34n TW35-49n TW50-64n TW65_AND_ABOVEn
DOCUMENT WEIGHT CALCULATION AGAINST GENDER PROFILES
D1 DWM1 DWF1
D2 DWM2 DWF2
Dm DWMn DWFn
DOCUMENT WEIGHT CALCULATION AGAINST AGE PROFILES
D1 DW18-241 DW25-341 DW35-491 DW50-641 DW65_AND_ABOVE1
D2 DW18-242 DW25-342 DW35-492 DW50-642 DW65_AND_ABOVE2
Dm DW18-24m DW25-34m DW35-49m DW50-64m DW65_AND_ABOVEm
DOCUMENT VECTORS FOR GENDER PREDICTION
<DWM, DWF, GENDER CLASS>
DOCUMENT VECTORS FOR AGE PREDICTION
<DW18-24, DW25-34, DW35-49, DW50-64, DW65 _AND_ABOVE, AGE CLASS>
TRAINING CLASSIFIER
TRAINING CLASSIFIER
CLASSIFICATION MODEL FOR GENDER PREDICTION
CLASSIFICATION MODEL FOR AGE PREDICTION
UNKNOWN DOCUMENT VECTOR
GENDER PREDICTION OF UNKNOWN DOCUMENT
AGE PREDICTION OF UNKNOWN DOCUMENT

Figure 1. The model for proposed document weighted approach

3. 1. Term Weights Specific To Profiles

Term weighting measures important concept in the modern information analysis. Different terms have different importance in a text. In general, Author profiling techniques analyze and predict the demographics of authors easily when the document contains large amount of text. For small texts it is difficult to predict the demographics features of the authors. As stated in literature [25] a normalized document length weight measure is used to assign suitable weights to terms in small sized texts. This weight measure as in Equation (1) is used the term frequency and unique terms in a document to find the term weight. The number of unique terms plays a significant 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 by increasing the size of the document.

Let $\{D_1,D_2,\ldots\ldots D_m\}$ is a collection of documents in the corpus, $P = \{p_1, p_2,\ldots\ldots p_q\}$ is the set of profiles, $V = \{t_1, t_2, \ldots\ldots t_n\}$ is a collection of vocabulary terms for analysis. Each term $t_i \in V$ is represented as a vector t_{ij}, i.e., $t_{ij} = \left(t_{i1}, t_{i2}, \ldots\ldots., t_{iq}\right)$, where the dimension t_{ij}

represents the term t_i weight specific to profile group p_j.

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

where, $W(t_i, p_j)$ is the weight of i^{th} term in j^{th} profile. TF_{ik} (Term Frequency) is the number of times the term t_i is occurred in a document k, $AVGTF_{ik}$ 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 term normalization [25], the constant value 0.2 is suitable for slope variable in Equation (1) for retrieval of suitable information. 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 terms that are occurred in more number of documents have more weight and the terms that occurred in less number of documents have less weight. Therefore, normalization of weight values is required. The Equations (2) and (3) are used to normalize the term weight values.

$$\overline{t_{ij}} = \frac{w(t_i, p_j)}{\sum_{i=1}^{n} w(t_i, p_j)} \tag{2}$$

$$t_{ij} = \frac{\overline{t_{ij}}}{\sum_{j=1}^{q} w(t_i, p_j)} \tag{3}$$

where, $\overline{t_{ij}}$ 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 $\overline{t_{ij}}$ to the weight of the term t_i in all the considered profiles.

3. 2. Document Weights Specific To Profiles

In this work, the document weight specific to profile group is computed by aggregating the term weights specific to the document and the term weights specific to the profile group. Equation (4) is used to calculate the weight of a document specific to a profile group. In Equation (4), Term Frequency Inverse Document Frequency (TFIDF) measure is used to calculate the weight of the term in a particular document. TFIDF measure as in Equation (5) assigns the weight to a term based on the term frequency and the number of documents contains the term in a corpus of profile group.

$$W_{d_{kj}} = \sum_{t_i \in d_k, d_k \in p_j} TFIDF(t_i, d_k) \cdot W_{t_{ij}} \tag{4}$$

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

$$TFIDF(t_i, d_k) = TF_{ik} * \log\left(\frac{|D|}{|1+DF_{ti}|}\right) \tag{5}$$

The document vectors are finally represented using Equation (6):

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

where, $z_k = \{W_{dk1}, W_{dk2}, ..., W_{dkq}\}$ and c_j is a class label of profile p_j. The vector Z contains document weights specific to each profile with document profile label. In this work, two profile groups are considered for gender such as male and female and five profile groups are considered for age such as 18-24, 25-34, 35-49, 50-64 and 65_AND_ABOVE.

In this approach, the outcome of the document vector representation for gender prediction is $\{Wd_{1male}, Wd_{1female}, C_1\}$. Where, Wd_{1male}, $Wd_{1female}$ are the document weights of d_1 in the male profile group and female profile group respectively, C_1 is a class label of a document d_1. For the age prediction, the outcome of the vector representation is $\{Wd_{118-24}, Wd_{125-34}, Wd_{135-49}, Wd_{150-64}, Wd_{165\_AND\_ABOVE}, C_1\}$. Where, Wd_{118-24}, Wd_{125-34}, Wd_{135-49}, Wd_{150-64}, $Wd_{165\_AND\_ABOVE}$ are the document weights of d_1 in the corpuses of 18-24, 25-34, 35-49, 50-64 and 65_AND_ABOVE age groups respectively.

The number of features used to represent a document depends on the number of profile groups in a profile. To avoid high dimensionality problem, the proposed approach used only two features for gender prediction and five features for age prediction to represent the document vectors. As in Equation (4), the document weight is computed by aggregating the term weights specific to the document and specific to profile group, thus the semantic relationship is captured between the terms in a document and corpus of documents.

4. ANALYSIS OF EXPERIMENTAL RESULTS

This experiment is carried out on hotel reviews using accuracy as a performance measure. Various machine learning classifiers such as Naive Bayes Multinomial, Simple Logistic, Logistic, IBK, Bagging and RandomForest are used from WEKA tool to generate the classification model. In WEKA tool, 10-fold cross validation is used to evaluate the set of document vectors. In 10-fold cross validation, the original corpus is randomly partitioned into 10 samples. Out of 10 samples, 9 samples are used for training classification model and the one is used for testing the performance of the classification model. This process is repeated until every sample is used exactly once as the validation data.

4. 1. Corpus Characteristics The corpus was collected from TripAdvisor.com, which contains 4000 reviews about different hotels. Table 1 shows the characteristics of the reviews corpus used for gender and age prediction. 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, the following steps are adopted. First, reviews containing less than five lines of text were excluded from our dataset. Second, the reviews are considered which are written in English language. Finally, the reviews were considered written by the authors whose gender and age information was given in their user profile.

After collecting the reviews two preprocessing steps were applied on the corpus such as stop words removal and stemming. In this experiment, it was adopted the stop word list as in web site[2] and the stemming is performed by using porter stemming algorithm [26]. 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.

4. 2. Evaluation Measures The researchers in 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 estimate the performance of the classification model. Accuracy as in Equation (4.1) is the ratio of the number of documents correctly predicted their profiles to the total number of documents in the corpus.

$$\text{Accuracy}=\frac{\text{Number of documents correctly predicted their profiles}}{\text{Total number of documents}} \qquad (4.1)$$

TABLE 1. Corpus characteristics

S No.	Age Group	Number of Reviews	Number of Male Reviews	Number of Female Reviews
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

4. 3. Identifying Author Profiles

4. 3.1. Gender Prediction The gender prediction is evaluated as a classification problem and accuracy measure was used to report the results. The results achieved for gender prediction in the proposed approach are presented in Table 2.

The proposed approach achieved good accuracies based on the effectiveness of the term weight measure and document weight measure. In Table 2, it was observed that as the number of terms increased from 1000 to 8000 with an interval of 1000 most frequent terms for computing the document weight, the growth rate in accuracy was increased. The Naïve Bayes Multinomial classifier achieved a best accuracy of 91.5% for gender prediction. It was also witnessed that, the accuracies are increased in all the classifiers when the number of terms for computing the document weight are increased.

Figure 2 shows the classifiers performance for gender prediction when the number of features are increased. The Naive Bayes Multinomial classifier obtained good accuracies than other classifiers when the number of features increased from 1000 to 8000. 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. The proposed approach used only two features for representing a document where as other approaches used more number of features to represent a document.

Table 3 represents the accuracies of existing approach and proposed approach for gender prediction when Naive Bayes Multinomila classifier is used on same reviews dataset. The existing approaches used the combination of stylistic features to differentiate the writing styles of the authors.

TABLE 2. The accuracy of gender prediction for various machine learning classifiers

Classifier/Number of Terms	NBM	SL	LOG	IBK	BAG	RF
1000	79.25	76.00	79.15	69.35	72.80	72.20
2000	82.15	79.05	81.95	73.60	74.20	75.35
3000	84.60	81.75	84.25	75.45	76.70	76.60
4000	86.35	83.95	86.30	79.75	78.55	79.60
5000	87.80	85.00	87.55	80.35	80.35	81.90
6000	89.75	87.60	89.05	82.85	81.90	83.65
7000	90.70	88.70	89.90	85.55	83.60	85.05
8000	**91.50**	90.00	90.85	85.80	84.55	86.50

[2] http://members.unine.ch/jacques.savoy/clef/index.html

The relationship between the features to profile is not captured in the existing approaches. In this approach, content based features such as most frequent terms are used to compute the document weight. Term weight measures are used to assign the discriminative power to the terms. In this work, a suitable term weight measure is identified to assign appropriate weights to the terms.

In general, every term is having a specific importance in different profile groups. For example ‘bowl’ is a term 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 term, the document which is written by male or female is not predicted certainly by using terms individually. This type of representation of a document with individual terms is not increased the predictive accuracy of profiles.

In this work, a new model is proposed to represent the document with document weights not with the weights of the terms in that document. In this model, the weight of bowl term is computed in all the documents of male and female documents. 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 the profile group. The proposed approach capture the relationship between terms by representing document vector with weights of the documents.

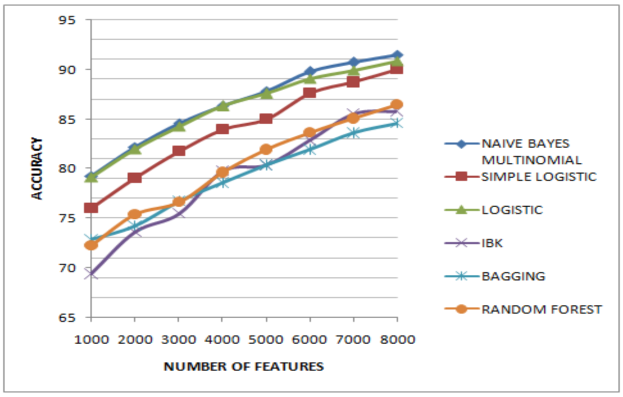

Figure 2. Performance of various classifiers for gender prediction

TABLE 3. Accuracies of proposed approach and existing approaches for gender prediction using naïve bayes multinomial classifier on reviews corpus

Approach	Number of Features	Accuracy
Estival D. et al. [21]	689	76.57
Dang Duc P. et al. [24]	298	79.91
Argamon S. et al. [6]	1000	81.26
J. Schler et al. [2]	1502	86.07
Proposed	2	**91.50**

The existing approaches extracted more number of features to represent a document and these features are independently participate in the classifcation process. In the proposed approach, the document weight is computed by using all the features in a specific document. All the features are collaboratively participate in the classification process. This is the reason that our proposed approach achieved good accuracies than existing approaches for gender and age prediction.

4. 3. 2. Age Prediction Table 4 shows the accuracies of age prediction for various classifiers. The logistic classifier achieved a highest accuracy of 81.58 % for age prediction among other classifiers by using 8000 most frequent terms to compute the document weight. 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. The accuracy is increased in all the classifiers when the number of terms is increased.

Figure 3 shows the comparison of classifiers performance for age prediction. The logistic classifier achieved good accuracies than other classifiers when the number of features is increased to compute the document weight.

Table 5 shows the comparisons of existing approaches with proposed approach for age prediction on same reviews corpus when logistic classifier is used. 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. Overall, the proposed approach achieved good accuracy for age prediction compared to the existing approaches on reviews corpus in Author profiling by using less number of features.

TABLE 4. The accuracy of age prediction for various machine learning classifiers

Classifier/Number of Terms	NBM	SL	LOG	IBK	BAG	RF
1000	47.07	34.23	54.15	31.35	33.80	34.81
2000	54.34	39.56	62.98	36.56	38.00	40.51
3000	59.24	44.35	67.33	43.67	41.99	45.11
4000	66.12	48.21	72.33	45.11	44.53	50.08
5000	69.87	52.38	74.99	49.25	47.13	52.77
6000	73.64	57.03	78.06	53.54	51.12	57.43
7000	76.25	59.85	79.89	55.99	52.47	61.30
8000	78.18	62.06	**81.58**	56.94	54.40	61.66

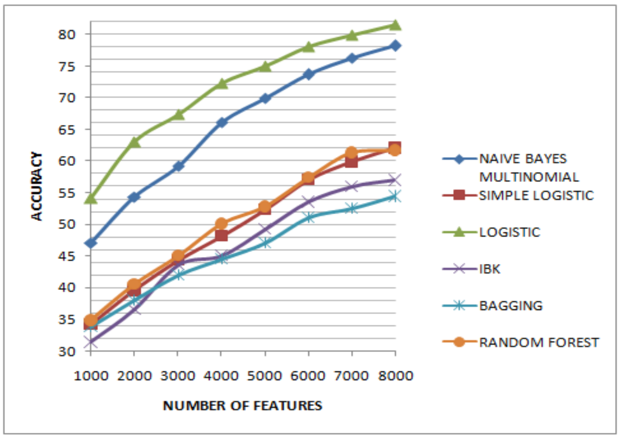

Figure 3. Performance of classifiers for age prediction

TABLE 5. The accuracies of proposed approach and existing approaches for age prediction on reviews corpus using logistic classifier

Approach	Number of Features	Accuracy
Estival D. et al. [21]	689	63.93
Dang Duc P. et al. [24]	298	73.02
Argamon S. et al. [6]	1000	79.67
J. Schler et al. [2]	1502	73.71
Proposed	2	**81.58**

5. CONCLUSIONS AND FUTURE SCOPE

In this paper, a new document weight approach 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. The proposed approach obtained an overall accuracy of 91.50% for gender prediction and 81.58 % for age prediction. This is the accuracy which is proved to be above the range of the accuracies achieved by the existing approaches with minimal number of features in Author profiling. The proposed approach used the most frequent terms to compute the document weight.

In our future work, it is planned to consider the domain characteristics and categoricl 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. In document weight computation the strenght of the term within a document is calculated with term frequency and inverse document frequency measure. It is also planned to replace inverse document frequency with inverse category frequency to compute the term weight within a document.

6. REFERENCES

1. Koppel, M., Argamon, S. and Shimoni, A.R., "Automatically categorizing written texts by author gender", ***Literary and Linguistic Computing***, Vol. 17, No. 4, (2002), 401-412.
2. Schler, J., Koppel, M., Argamon, S. and Pennebaker, J.W., "Effects of age and gender on blogging", in AAAI spring symposium: Computational approaches to analyzing weblogs. Vol. 6, (2006), 199-205.
3. Nerbonne, J., *The secret life of pronouns. What our words say about us*. 2013, ALLC.
4. Newman, M.L., Groom, C.J., Handelman, L.D. and Pennebaker, J.W., "Gender differences in language use: An analysis of 14,000 text samples", ***Discourse Processes***, Vol. 45, No. 3, (2008), 211-236.
5. Pennebaker, J.W., Francis, M.E. and Booth, R.J., "Linguistic inquiry and word count: Liwc 2001", ***Mahway: Lawrence Erlbaum Associates***, Vol. 71, No. 2001, (2001), 2001-2009.
6. Argamon, S., Koppel, M., Pennebaker, J.W. and Schler, J., "Mining the blogosphere: Age, gender and the varieties of self-expression", ***First Monday***, Vol. 12, No. 9, (2007).
7. Santosh, K., Bansal, R., Shekhar, M. and Varma, V., "Author profiling: Predicting age and gender from blogs", ***Notebook for PAN at CLEF***, (2013), 119-124.
8. Argamon, S., Koppel, M., Pennebaker, J.W. and Schler, J., "Automatically profiling the author of an anonymous text", ***Communications of the ACM***, Vol. 52, No. 2, (2009), 119-123.
9. Sapkota, U., Solorio, T., Montes-y-Gomez, M. and Ramírez-de-la-Rosa, G., "Author profiling for english and spanish text", ***Notebook for PAN at CLEF***, Vol., No., (2013).
10. Hamidi, H. and Daraee, A., "Analysis of pre-processing and post-processing methods and using data mining to diagnose heart diseases", ***International Journal of Engineering-Transactions A: Basics***, Vol. 29, No. 7, (2016), 921-929.
11. Lim, W.-Y., Goh, J. and Thing, V.L., "Content-centric age and gender profiling", ***Proceedings of the Notebook for PAN at CLEF***, (2013), 130-138.
12. Darvishi, A. and Hassanpour, H., "A geometric view of similarity measures in data mining", ***International Journal of Engineering-Transactions C: Aspects***, Vol. 28, No. 12, (2015), 1728-1735.
13. Mechti, S., Jaoua, M., Belguith, L.H. and Faiz, R., "Author profiling using style-based features", in Proceedings of CLEF, Citeseer., (2013).
14. Maharjan, S., Shrestha, P. and Solorio, T., "A simple approach to author profiling in mapreduce", in CLEF (Working Notes)., (2014), 1121-1128.
15. Grivas, A., Krithara, A. and Giannakopoulos, G., "Author profiling using stylometric and structural feature groupings", in CLEF (Working Notes)., (2015).
16. Palomino-Garibay, A., Camacho-Gonzalez, A.T., Fierro-Villaneda, R.A., Hernandez-Farias, I., Buscaldi, D. and Meza-Ruiz, I.V., "A random forest approach for authorship profiling", ***Cappellato et al.[8]***, (2015), 156-164.
17. Octavia-Maria, S., "Ulea1; 2 and daniel dichiu, bitdefender romania,"automatic profiling of twitter users based on their tweets."", in Proceedings of CLEF., (2015).
18. Weren, E.R., Moreira, V.P. and de Oliveira, J.P.M., "Exploring information retrieval features for author profiling", in CLEF (Working Notes)., (2014), 1164-1171.
19. Weren, E.R., Moreira, V.P. and Oliveira, J., "Using simple content features for the author profiling task", in Notebook for PAN at Cross-Language Evaluation Forum. Valencia, Spain., (2013).

20. Weren, E.R.D., "Information retrieval features for personality traits", in CLEF (Working Notes)., (2015).

21. Estival, D., Gaustad, T., Pham, S.B., Radford, W. and Hutchinson, B., "Author profiling for english emails", in Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING'07)., (2007), 263-272.

22. Soler, J. and Wanner, L., "How to use less features and reach better performance in author gender identification", in LREC., (2014), 1315-1319.

23. Pham, D.D., Tran, G.B. and Pham, S.B., "Author profiling for vietnamese blogs", in Asian Language Processing, 2009. IALP'09. International Conference on, IEEE., (2009), 190-194.

24. Dang, D., Giang, B. and Bao, P., "Authorship attribution and gender identification in greek blogs", in 8th International Conference on Quantitative Linguistics (QUALICO)., (2012), 26-29.

25. Singhal, A., Buckley, C. and Mitra, M., "Pivoted document length normalization", in Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, ACM., (1996), 21-29.

26. Porter, M., "Developing the english stemmer", http://snowball.tartarus.org, (2002).

A Document Weighted Approach for Gender and Age Prediction Based on Term Weight Measure

T. Raghunadha Reddy[a], B. Vishnu Vardhan[b], P. Vijayapal Reddy[c]

[a] Department of Information Technology,Vardhaman College of Engineering, Hyderabad, Telangana, India
[b] Department of Computer Science and Engineering, JNTUH College of Engineering, Jagtiyal, Karimnagar, Telangana, India
[c] Department of Computer Science and Engineering, Matrusri Engineering college, Hyderabad, Telangana, India

PAPER INFO

Paper history:
Received 02 September 2016
Received in revised form 11 February 2017
Accepted 10 March 2017

Keywords:
Author Profiling
Gender Prediction
Age Prediction
Term Weight
Document Weight
Classifiers

چکیده

پروفایل کردن نویسنده یک روش طبقه بندی متن است که برای پیش بینی پروفیل های متن ناشناخته با تجزیه و تحلیل سبک نوشتن آنها استفاده شده است. پروفایل های نویسنده، ویژگی های نویسندگان مانند جنس، سن، زبان بومی، کشور و پس زمینه آموزشی می باشد .روش های موجود برای پروفایل کردن نویسنده از مشکلاتی مانند ابعاد بالا ویژگی های رنج می برد و از به تصرف درآوردن رابطه بین ویژگی ها شکست می خورد. در این کار، یک رویکرد بر پایه سند جدید به منظور رسیدگی به مشکلات در روش های موجود ارائه شده است .در این روش، واژه اندازه گیری وزن برای اختصاص دادن وزن مناسب به شرایط استفاده شده است و این وزن های واژه برای محاسبه وزن سند جمع آوری شده است. مدل طبقه بندی با این وزن های سند برای پیش بینی پروفیل هایی از متن تولید شده است .روش پیشنهادی و روش های موجود در بررسی دامنه با طبقه بندی های مختلف تجربه شده اند. توجه و دقت روش پیشنهادی برای جنسیت و پیش بینی سن امیدوار کننده تر از روش های موجود می باشد.

doi: *10.5829/idosi.ije.2017.30.05b.03*