Advances in Computational Sciences and Technology
ISSN 0973-6107 Volume 10, Number 7 (2017) pp. 2061-2068

http://www.ripublication.com

Impact of Term Weight Measures in Author Profiling Approaches for Gender Prediction

G Srikanth Reddy[1] and T Raghunadha Reddy[2]

[1] *Dept of IT, Vardhaman College of Engineering, Hyderabad, India.*

Abstract

In recent times the text in the internet is growing exponentially through bogs, social media, twitter tweets and reviews. Most of the text generated by the authors in the internet is anonymous. Author profiling is a text classification technique which is used to predict the demographic profiles such as gender, age, native language, location and educational background of the anonymous text by analyzing their writing styles. The researchers proposed various features such as character based, word based, structural, syntactic and semantic features to differentiate the writing styles of the authors. The existing approaches used the frequency of a feature to represent the document vector. In this work, we experimented with various features with their frequency and observed that only frequency is not suitable to assign better discriminative power to the features. Later, we analyzed various term weight measures from various domains to assign suitable weights to the terms. In this approach, we experimented on reviews domain and obtained good accuracy for gender prediction when compared to existing approaches.

Keywords: Author Prediction, Gender Prediction, Term Weight Measure, BOW approach.

1 INTRODUCTION

The internet is growing rapidly with textual information along with the cyber crimes also increased in the internet. The people are sending harassing messages in social media and the terrorist organizations send threatening mails without specifying their correct details. The researchers are attracted to know the demographic characteristics

of these texts by analyzing the writing styles. Author Profiling is one such area to predict the profiling characteristics of the authors. Author profiling is used in several information process enabled applications such as forensic analysis, marketing, educational domain, security and psychology.

In general, every human has his own writing style and he continues his own style while writing in blogs, social media, twitter tweets and reviews. According to Koppel et al. [1], female use more number of pronouns and stress more on topics related to shopping, beauty and kitty parties in their writings. Male write more about topics related to technology, politics and sports in their writings and use more number of quantifiers and determiners than females. J. Schler et al., [2] observed that the male authors use more prepositions than female in their articles and blog posts. In another observation [3], it is informed that females use more adjectives and adverbs than male authors and write more about wedding styles in their writings.

The main focus of this paper is to predict the gender of the authors in reviews domain. This paper is organized as follows: section 2 explains the related work in Author Profiling. The existing approach used in Author Profiling is described in section 3. Various term weight measures from different domains are discussed in section 4. The section 5 analyzes the accuracies of gender prediction. Section 6 concludes this work suggested future directions.

2. RELATED WORK

The general approach followed in most of the existing approaches for Author Profiling was BOW (Bag Of Words) approach. In Author Profiling, the main concentration of the researchers is to extract the features that are suitable for differentiating the writing styles of the authors. Various researchers proposed different types of features to represent the documents [4].

In Koppel, M. et al. [1], 566 documents are collected from the British National Corpus (BNC). They extracted 1081 features and achieved an accuracy of 80% for gender prediction using exponential gradient algorithm. In another work [2], they achieved better accuracy of 80.1% using multi class real winnow algorithm with 1502 features on 37478 blogs corpus. Estival D. et al. experimented [5] on emails dataset and extracted 689 features. Different machine learning algorithms were applied on the dataset and observed that SMO obtained best accuracy 69.26% for gender prediction compared to other classifiers.

Argamon, S. et al. [6] experimented on blog posts of 19320 blog authors. It was observed that the style based features were most useful to discriminate the gender and bayesian multinomial regression classifier obtained an accuracy of 76.1% for gender prediction. Juan Soler Company et al. worked [7] on the corpus of New York Times

opinion blogs. They tried with 83 features for gender prediction and achieved good accuracy of 82.83% using Bagging classifier.

3. TRADITIONAL APPROACH

3.1 Bag Of Words (BOW) Approach

Fig. 1 represents the architecture of the BOW approach. In this approach, first the preprocessing techniques are applied on the collected dataset. Extract the most frequent terms or features that are important to discriminate the writing styles of the authors from the modified dataset. Consider these terms or features as bag of words. Every document in the dataset is represented with this bag of words. Each value in the document vector is the weights of the bag of words. Finally, the document vectors are used to generate classification model.

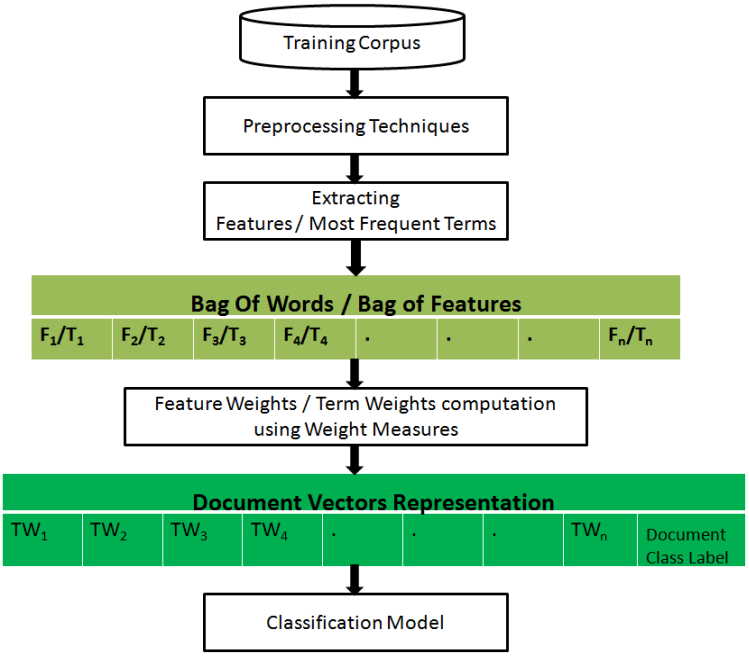

Fig. 1. The Model of BOW Approach

3.2 Dataset Characteristics

The dataset was collected from tripadvisor.com and it contains 4000 english reviews about different hotels. In order to make this dataset applicable to Author Profiling and to ensure its quality, the reviews were considered written by the authors whose gender was given in their user profile.

3.3 Evaluation Measures

In this work, accuracy measure is used to evaluate the performance of the classifier. Accuracy is the ratio of number of test documents correctly predicted their gender and the number of test documents considered. Accuracy measure is represented as

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

4. TERM WEIGHT MEASURES

Traditional term weighting measures are Binary, Term Frequency (TF) and Term Frequency Inverse Document Frequency (TFIDF). Although the TFIDF has been proved in Information Retrieval domain and several text mining tasks for quantifying the term weights, but it is not most effective for Author Profiling because TFIDF disregard the profile group information of the training documents. Therefore researchers are looking for alternative effective term weight measures in Author Profiling.

The term weight measures are categorized into two types such as unsupervised and supervised term weight measures based on the usage of the class label information. An unsupervised term weight measure does not use information regarding class label. The supervised term weight measure use class label information.

4.1 Unsupervised Term Weight Measures

4.1.1 Nonuniform Distributed Term Weight (NDTW) Measure

The main principle of NDTW measure is assigning more weight to the terms that are not uniformly distributed across the documents [8]. NDTW measure is represented in equation (1).

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

$TOTF_{ti}$ computes the total occurrence of the term t_i in profile p_j and captures intra-class distribution of term t_i, $tf(t_i, d_k)$ calculates number of times term t_i occurred in document d_k and captures the inner-class distribution of term t_i.

4.1.2 Normalized Document Length Term Weight (NDLTW) Measure

In this work, a Normalized Document Length Term Weight (NDLTW) Measure is used to analyze small sized texts [9]. The NDLTW measure is represented in equation (2).

$$W_{tij} = W(t_i, p_j) = \sum_{k=1}^{m} \frac{(1+\log(TF_i)) / (1+\log(AVGTF_i))}{(1-slope) * AVGUT_k + slope * UT_k} \quad (2)$$

$W(t_i, p_j)$ is the weight of i^{th} term in j^{th} profile. TF_i (Term Frequency) is the frequency of the term t_i in document d_k, $AVGTF_i$ is a ratio of the term frequency t_i to the total number of terms in k^{th} document and slope value is 0.2. UT_k is a number of unique terms in d_k document, and $AVGUT_k$ is a ratio of unique terms to total number of terms in k^{th} document.

4.2 Supervised Term Weight Measures

4.2.1 Relevance Frequency based Term Weight (RFTW) Measure

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 [10]. The RFTW measure is shown in equation (3).

$$tf * rf = tf * \log\left(2 + \frac{a}{\max(1, c)}\right) \quad (3)$$

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

4.2.2 Inverse Query and Category Frequency Term Weight (IQCFTW) Measure

This weight measure as shown in equation (4) was proposed in [11]. *qf* (question frequency) of term t_i is the number of positive documents contains the term t_i. *cf* (category frequency) of a term t_i is the number of categories in which t_i occurs. icf (inverse category frequency) is the ratio of total number of categories and the category frequency of a term. iqf (inverse question frequency) similar to IDF.

$$iqf * qf * icf = \log\left(\frac{N}{tp + fn}\right) * \log(tp + 1) * \log\left(\frac{|C|}{cf} + 1\right) \quad (4)$$

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 that contains the term ti, $|C|$ is the number of categories.

4.2.3 Discriminative Feature Selection Term Weight (DFSTW) Measure

DFS measure assigns more weight to the terms that are having high average term frequency in class c_j and the terms with high occurrence rate in most of the documents of c_j [12]. The DFSTW measure is represented in equation (5).

$$W(t_i, c_j) = \frac{tf(t_i, c_j) / df(t_i, c_j)}{tf(t_i, \overline{c_j}) / df(t_i, \overline{c_j})} \times \frac{a_{ij}}{(a_{ij} + b_{ij})} \times \frac{a_{ij}}{(a_{ij} + c_{ij})} \times \left| \frac{a_{ij}}{(a_{ij} + b_{ij})} - \frac{c_{ij}}{(c_{ij} + d_{ij})} \right| \quad (5)$$

Where, a_{ij} and b_{ij} is the number of documents of class c_j which contain and which does not contain term t_i respectively. c_{ij} and d_{ij} is the number of documents which contain and which do not contain term t_i that do not belong to class c_j respectively.

5. EMPIRICAL EVALUATIONS

5.1 Results of Term Weight Measures

The performance of seven term weight measures on BOW approach is depicted in Fig. 2.

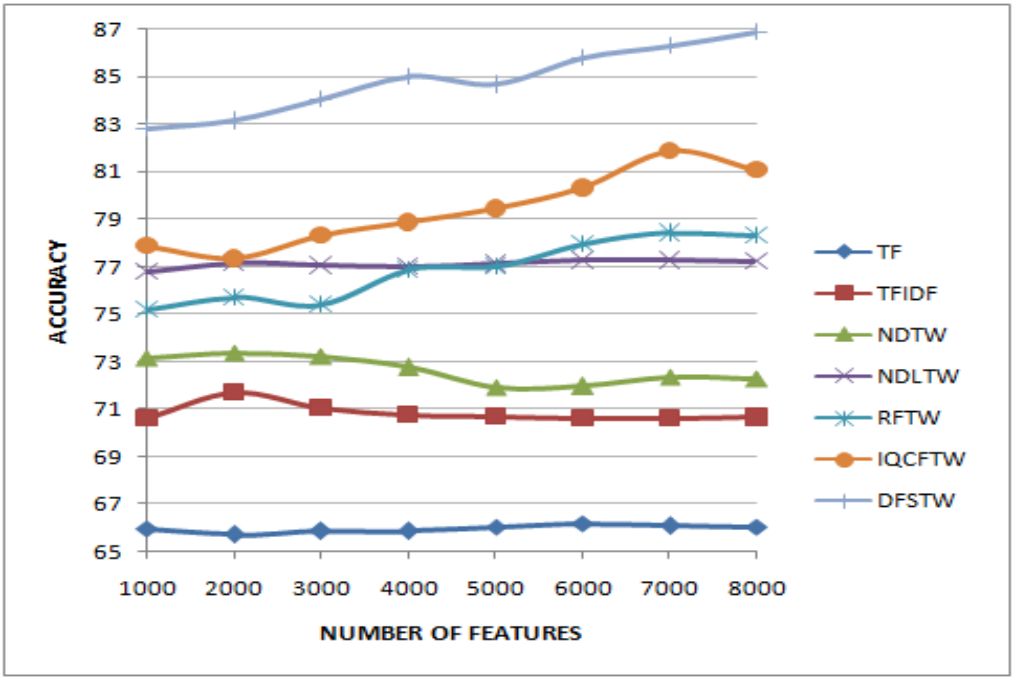

Fig. 2. Performance of term weight measures for gender prediction using Naïve Bayes Multinomial Classifier

The DFSTW measure produces accuracies that are superior to the other six term weight measures over the full range of various numbers of features in BOW approach. The DFSTW measure shows accuracy reduction when the number of features changes from 4000 to 5000. DFSTW measure obtained an accuracy of 86.91%, it is far better than the accuracy of other term weight measures for gender prediction using Naïve Bayes Multinomial classifier.

6. CONCLUSIONS AND FUTURE SCOPE

In this work, the experimentation is carried out various term weight measures. The DFSTW measure obtained good accuracy of 86.91% for gender prediction when compared with other term weight measures. In our future work, it is planned to predict other demographic profiles of the authors and also planned that propose a new term weight measure to increase the prediction accuracy of author profiles.

REFERENCES

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

[2] J. Schler, Moshe Koppel, S. Argamon and J. Pennebaker (2006), Effects of Age and Gender on Blogging, in Proc. of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, March 2006.

[3] Pennebaker, J.: The Secret Life of Pronouns: What Our Words Say About Us. Bloomsbury USA (2013), http://books.google.co.in/books?id=mJ4tLwEACAAJ

[4] 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.

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

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

[7] Juan Soler Company, Leo 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).

[8] Dennis, S.F., "The Design and Testing of a Fully Automated Indexing-Searching System for Documents Consisting of Expository Text",in Informational Retrieval: A Critical review, g. Schecter, editor,Thompson Book Company, Washington D.C., 1967, p.p 67-94.

[9] 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, pp. 21–29.

[10] 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.

[11] Xiaojun Quan, Wenyin Liu, and Bite Qiu, "Term Weighting Schemes for Question Categorization", IEEE transactions on pattern analysis and machine intelligence, vol. 33, no. 5, pp. 1009-1021, may 2011.

[12] Wei Zong , Feng Wu, Lap-Keung Chu, Domenic Sculli, "A discriminative and semantic feature selection method for text categorization", International Journal of production Economics, Elsevier, Jan 2015, pp. 215-222.