NATIVITY LANGUAGE PREDICTION USING A DOCUMENT WEIGHTED APPROACH

Raghunadha Reddy. T[1], P. Vijayapal Reddy[2]
[1]*Associate Professor, Dept of IT, Vardhaman College of Engineering, Hyderabad*
[2]*Professor, Department of CSE, Matrusri Engineering College, Saidabad, Hyderabad,*

Abstract: The task of analyzing an interested text to find demographic characteristics such as gender, age, country and nativity language of an unknown author based on the writing style of the text called Author Profiling. The writing style differences are very important in Author Profiling research. Different researchers proposed several stylistic features to differentiate the writing style of the authors. In this work, the experimentation carried out with content based features like most frequent terms in the corpus to predict the nativity language of the authors. A Profile specific Document Weighted approach is used with most frequent terms as features. In this approach, the document weights are used to represent the document vectors. Various classification algorithms were used to generate the classification model and to predict the nativity language of the authors. The achieved results were good when compared with exiting techniques for nativity language prediction in Author Profiling research area.

Index Terms: Author Profiling, Profile specific Document Weighted approach, Nativity Language Prediction, Term Weight Measure, Document Weight Measure.

I. INTRODUCTION

Author profiling is a text classification technique, which used to create a profile of an author of a text. Such a profile can include the gender, age, location, native language and the personality traits of the author. In author profiling, linguistic features are used to determine the profile of an author and the most common techniques that are used are different kind of machine learning techniques.

Author profiling is an important task in many different domains. For instance, from a marketing perspective, companies may be interested in knowing more about anonymous reviewers written on various product review sites. In forensic linguistics, author profiling can be used to determine the linguistic profile of an author of a suspicious text. This is something that can be valuable for evaluating suspects and as a support in investigations.

From an intelligence perspective author profiling is used to gain more information about a possible suspect - this could for example be a potential violent lone actor that reveals an intention of committing targeted violence in a online setting, something that research has shown is the case in previous attacks. Author Profiling is used to extract as much information as possible about an author may increase law enforcement's chances to advance in their investigations.

Machine learning is the most common approach for author profiling. However, there are some challenges when using these kind of techniques in realistic scenarios. First of all, the availability of labeled data that is used to train machine learning models on is limited. Secondly, the machine learning models may work well on the domain that they are trained and tested on but it is very difficult to predict how they will work on other domains. Therefore, it is very difficult to estimate the accuracy of the results when using author profiling in real-life scenarios. An important note here is to always involve human analysts in the usage of author profiling and only consider this kind of techniques as a support for humans in their analysis.

This paper is organized in 6 sections. The existing work in Author Profiling for nativity language prediction was explained in section 2. The corpus characteristics and performance evaluation measures were explained in section 3. The Profile specific Document Weighted model was described in section 4. The experimental results of PDW model represented in section 5. Section 6 concludes this paper with conclusions and future work.

II. RELATED WORK

Most of the researchers in Author Profiling proposed different types of stylistic features such as character, word, syntactic, content, structural, readability and information retrieval features for predicting the demographic characteristics of the authors[1]. Shlomo Argamon used [2] the corpus of International Corpus of Learner English (ICLE) which is a culmination of non-native English speakers from various countries and the corpus was tested to predict the age, gender and native language. He also used essays of 251 psychology undergraduates at the University of Texas at Austin for neurotism prediction. They considered five sub-corpora namely Russian, Czech Republic, Bulgaria, French and Spanish from ICLE. They used 258 authors writings from each sub corpus to avoid class imbalance problems. While predicting the age, gender, nativity language and neurotism they observed that style based features gave an accuracy of 65.1%, content based features gave an accuracy of 0.823 and both style based and content based features together gave an accuracy of 0.793.

Adamearcia et al., proposed [3] a simple classification method based on similarity among objects. They considered various features such as lemmas and grammatical category of lexical terms, tweets characteristics, subjectivity features and opinion mining analysis for document representation. They predicted gender and language variety from English language tweets. It was observed that they got good results for gender prediction and low accuracy for language variety prediction. Basile et al., experimented [4] with a single model to predict gender and language variety of four language texts. They used linear support vector machines with a set of features like character 3 to 5 grams, word unigrams,

POS tags, twitter handles and geographic entities for generating classification model. They obtained an accuracy of 82.53% for gender prediction and 91.84% for language variety prediction.

Martinc et al., proposed [5] a logistic regression classifier and experimented with different types of features such as character n-grams, word unigrams, word bigrams, word bound character tetragrams, punctuation trigrams, suffix character tetragrams, POS trigrams, character flooding count, document sentiment information, Emojis counts, wordlists specific to language variety. They experimented with different classifiers like Linear SVM, Logistic Regression, Random Forest and XGBoost. Among these classifiers Logistic Regression classifier obtained good accuracy of 86.63% for language variety prediction in English language. They got good accuracies in Portuguese language for gender and language variety prediction.

Ciobanu et al., used [6] character and word n-grams as features and multiple linear SVM as classifier for predicting gender and language variety of authors in different languages. they obtained 75% accuracy for language variety prediction in English language. ogaltsov et al., experimented [7] with high order character n-grams as features and Logistic Regression as classifier to predict gender and language variety of authors in different languages. they achieved 80.92% accuracy for language variety prediction in English language.

III. CORPUS CHARACTERISTICS AND EVALUATION MEASURES

In this section, the corpus characteristics and the measures for evaluating the performance of a classification model is discussed.

3.1 Dataset characteristics

In this work, PAN17-twitter corpus was used for experimentation which was released in 2017. This corpus consists of Twitter posts labeled with gender and their specific variation of their native language English (Australia (AS), Canada (CN), Great Britain (GB), Ireland (IL), New Zealand (NZ), United States (US)). The distribution of the classes is balanced. In total, PAN17-twitter consisted of 360 000 posts with a distribution of 60 000 posts for each class. Table 1 shows the characteristics of the corpus.

Table 1: The characteristics of PAN 17 Twitter English corpus for Nativity Language prediction

Number of Authors	Name	Labels	Number of posts	Label distribution	
3600	PAN17-twitter	Native Language	360000	Ireland	60000
				Canada	60000
				Great Britain	60000
				New Zealand	60000
				United States	60000
				Australia	60000

3.2 Performance measures

In general, when evaluating the results in the experiments in Author Profiling approaches the researchers used accuracy, precision, recall and F1-score was taken into consideration. Accuracy is the most commonly used performance measure which measures the proportion of all predictions that are correct. Using classification accuracy alone for evaluating the performance of the classification algorithm could be misleading, especially if the dataset is unbalanced or contains more than two classes. In this work, accuracy measure was used to evaluate the performance of the classifiers.

IV. PROFILE SPECIFIC DOCUMENT WEIGHTED APPROACH

The Profile specific Document Weighted (PDW) approach is proposed in [8]. The model of PDW approach is depicted in fig 1. In this approach, first the English corpus for nativity language prediction was collected from PAN 2017 competition. Then, preprocessing techniques were applied on the corpus to prepare the content for further analysis. The most frequent terms were extracted from the updated corpus. The term weight measure is used to compute the weights of the terms specific to each nativity language country of documents. The document weight measure is used to calculate the document weight specific to each nativity language country by using the weights of the terms in that document. The document vectors were represented with these document weights. Finally, these document vectors were given to different classification algorithms to generate the classification model and this model is used to predict the nativity language of unknown document.

In this approach, finding appropriate term weight measure and document weight measures are important to improve the accuracy prediction of nativity language prediction.

4.1 Term Weight Measure

The term weight measures assign suitable weights to the terms by considering different types of terms distribution information such inner-document (term distribution within a document), intra-class (term distribution within a positive class of documents) and inter-class distribution (term distribution across classes of documents) in the corpus of documents. In this work, a Supervised Unique Term Weight (SUTW) measure [9] is used to find the weight of the terms specific to every nativity language country. The SUTW measure is shown in equation (1).

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

In this measure, $tf(t_i,d_k)$ is the number of times term t_i occurred in document d_k, $tf(t_i,p_j)$ is the number of times term t_i occurred in p_j, d_{tk} is the number of terms in document d_k, UT_k is the number of terms that occurred once in document d_k, $AVGUT_k$ is the average number of unique terms in document d_k. a_{ij}, b_{ij} is the number of documents in profile p_j which contain the term t_i and which does not contain term t_i respectively. c_{ij}, d_{ij} are the number of documents of other than profile p_j which contain the term t_i and which does not contain the term t_i respectively.

4.2 Document Weight Measure

The document weight measure computes the document by using the weights of the terms in that document. In this work, a document weight measure [10] is used to compute the document weight. Equation (2) shows the document weight d_k specific to profile p_j.

$$W_{dkj} = \sum_{t_i \in d_k, d_k \in p_j} TFIDF(t_i, d_k) \cdot W_{tij} \quad (2)$$

Where, TFIDF(ti,dk) is the term frequency and inverse document frequency of term ti in document dk and it is represented in equation (3). TFIDF measure assigns more weight to the terms which are occurred in less number of documents. W_{tij} is the weight of the term ti in Profile pj.

$$TFIDF(t_i, d_k) = tf(t_i, d_k) * \log\left(\frac{|D|}{|1 + DF_{ti}|} \right) \quad (3)$$

Where, |D| is the number of documents in the corpus, DFti is the number of documents which contain the term ti.

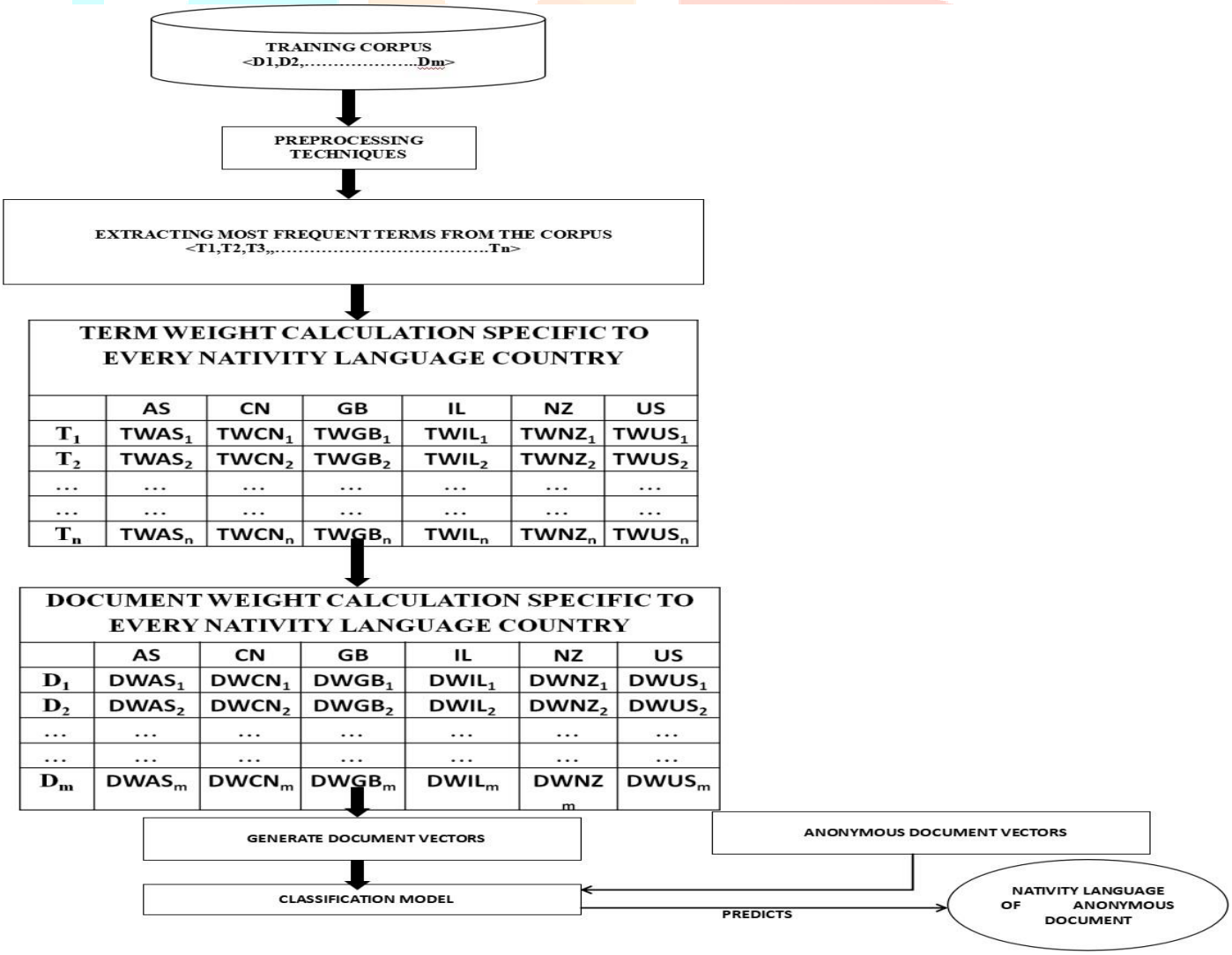

Fig. 1. *The model of Profile specific Document Weighted approach*

V. EXPERIMENTAL RESULTS

In this work, 8000 most frequent terms were extracted from the corpus. The experiment starts with 1000 terms and increased up to 8000 terms by adding 1000 terms in each iteration. By analyzing the corpus of language variety profile, it was observed that the content based features like the terms they used in their writings are different for different nativity language countries authors. It was understand that the selection of words to write a review is almost same for the users of one nativity language country. With this assumption most frequent 8000 terms were extracted from the corpus as features. Different classification algorithms such as Simple Logistic (SL), Logistic (LOG), Bagging (BAG), IBK, Naïve Bayes Multinomial (NBM) and Random Forest (RF) were used to generate the classification model. We obtained good results for nativity language prediction when compared with existing approaches of Author Profiling for nativity language prediction.

Table 2: The Accuracy of PDW model for nativity language of authors prediction using different Classifiers

Classifiers/ Number of Terms	SL	IBK	LOG	BAG	NBM	RF
1000	65.93	61.72	70.28	65.77	70.07	71.92
2000	66.14	62.57	71.11	68.59	74.54	75.89
3000	69.62	64.39	74.42	70.31	77.42	78.34
4000	71.38	65.74	75.20	71.72	78.21	81.71
5000	72.15	68.83	77.78	72.84	79.78	83.47
6000	72.08	70.24	78.46	74.27	81.46	85.41
7000	74.93	72.18	80.52	75.13	82.52	86.91
8000	75.74	73.92	81.14	77.22	84.63	88.57

The accuracies of nativity language prediction in PDW model using term weight measures are shown in Table 2. The Random Forest classifier obtained highest accuracy of 88.57% for nativity language prediction when compared with all other classifiers. The Naïve Bayes Multinomial classifier obtained an accuracy of 84.63% for nativity language prediction. It was observed that in all classifiers the accuracies of nativity language prediction was increased when the number of terms increases.

VI. CONCLUSION AND FUTURE SCOPE

This work was delimited to author profiling of native language on text written in English language. In addition, it was observed that how the PDW approach performed on different classifiers by comparing a number of different machine learning models for author profiling. Out of six evaluated classifiers, the results showed that the overall best performing classifier was the Random Forest classifier which outperformed the other classifiers. The Random Forest classifier obtained 88.57% accuracy for nativity language prediction when 8000 terms were used as features. How well each model performs on author profiling depends on many factors such as the size of the dataset, the balancing of the dataset is and the preprocessing techniques used to prepare the data for classification.

A potential extension would be to perform author profiling on text in other languages such as Spanish, Dutch and Arabic of PAN 2017 competition using similar models as used in this work. Also, another potential extension could be to discern information as personality traits, such information were for example included in the PAN 2015 dataset.

REFERENCES

[1] T. Raghunadha Reddy, B.VishnuVardhan, and P.Vijaypal Reddy, "A Survey on Authorship Profiling Techniques", International Journal of Applied Engineering Research, Volume 11, Number 5 (2016), pp 3092-3102.
[2] 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):119.(2009).
[3] Yaritza Adame-Arcia, Daniel Castro-Castro, Reynier Ortega Bueno, Rafael Mu-ñoz, " Author Profiling, instance-based Similarity Classification", Proceedings of CLEF 2017 Evaluation Labs, 2017.
[4] Angelo Basile, Gareth Dwyer, Maria Medvedeva, Josine Rawee, Hessel Haagsma, and Malvina Nissim, " N-GrAM: New Groningen Author-profiling Model Notebook for PAN at CLEF 2017", Proceedings of CLEF 2017 Evaluation Labs, 2017.
[5] Matej Martinc, Iza Škrjanec, Katja Zupan, and Senja Pollak, " PAN 2017: Author Profiling - Gender and Language Variety Prediction", Proceedings of CLEF 2017 Evaluation Labs, 2017.
[6] Alina Maria Ciobanu, Marcos Zampieri, Shervin Malmasi, Liviu P. Dinu, " Including Dialects and Language Varieties in Author Profiling", Proceedings of CLEF 2017 Evaluation Labs, 2017.
[7] Alexander Ogaltsov and Alexey Romanov, " Language Variety and Gender Classification for Author Profiling in PAN 2017", Proceedings of CLEF 2017 Evaluation Labs, 2017.
[8] Raghunadha Reddy T, Vishnu Vardhan B, Vijayapal Reddy P, "Profile specific Document Weighted approach using a New Term Weighting Measure for Author Profiling ", International Journal of Intelligent Engineering and Systems, Nov 2016, 9 (4), pp. 136-146. DOI: 10.22266/ijies2016.1231.15
[9] Raghunadha Reddy T, Vishnu Vardhan B, Vijayapal Reddy P, "A Document weighted Approach for Gender and Age Prediction", International Journal of Engineering, Volume 30, No. 5, 2017, PP. 647-653.
[10]Raghunadha Reddy T, Vishnu Vardhan B, Vijayapal Reddy P, "Author profile prediction using pivoted unique term normalization", Indian Journal of Science and Technology, Vol 9, Issue 46, Dec 2016.