Experimental Study Of Gender And Language Variety Identification in Social Media

Vineetha Rebecca Chacko, Anand Kumar M, Soman K P.

Abstract Social media has evolved to be a crucial part of life today for everyone. With such a global population communicating with each other, comes the accumulation of large amounts of social media data. This data, can be categorized as "Big Data", owing to its large quantity. It contains valuable information in the form of the demographics of authors on online platforms, the analysis of which is required in certain scenarios to maintain decorum in the online community. Here, we have analyzed Twitter data, which is the training data of the PAN@CLEF 2017 shared task contest, to identify the gender, as well as the language variety of the author. It is available in 4 different languages, namely, English, Spanish, Portuguese and Arabic. Both Document Term Matrix (DTM) and Term Frequency - Inverse Document Frequency (TF-IDF) have been used for text representation. The classifiers used are SVM, AdaBoost, Decision Tree and Random Forest.

1 Introduction

"Data is the new gold" - When it comes to something as valuable as gold, and is available in abundance, its handling also gets strenuous. Where traditional data handling techniques failed to tame such large amounts of data, there came up the field of **Big Data**, specialized in handling large volumes of complex data. With its different dimensions like veracity and variety, several challenges like storage, searching through the data, sharing of data, efficient privacy protection and most importantly, the analysis of data are faced. Analysis is the key ingredient of almost

Vineetha Rebecca Chacko, Anand Kumar M, Soman K. P.
Centre for Computational Engineering and Networking (CEN), Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, India.
e-mail: cb.en.p2cen16017@cb.students.amrita.edu, m_anandkumar@cb.amrita.edu, kp_soman@amrita.edu.

all fields of study, that proves, or disproves, all initial hypothesis made by man, about any scenario at hand.

Social media is a concept that came into limelight in the recent years. Indians got introduced to it all, mostly through platforms like Google Talk, Orkut etc. which gained popularity at an exponential rate, only to be followed by other social media giants like Facebook, Twitter and WhatsApp. Along with its advantages, the biggest of which is real time communication, it has certain negative aspects, like cyber stalking, cyber bullying, hacking, and even the spread of fake news. One of the biggest examples of our times is the investigation on the 2016 US election, which was claimed to be rigged by the Russians; and social media played a huge role in it. Summing up all these negative aspects of social media, the aspect that serves as a common helping hand in achieving these heinous goals, is **anonymity**. Anonymity is a treacherous enemy to the decorum of any online community.

Since everybody has access to the internet (except for the North Koreans), people have the freedom to hide behind fake profiles on social media platforms. Hence came up the field of cyber forensics facing the challenge of **Author Profiling**. It is the analysis of the demographics, like the nativity, gender etc of online authors. PAN@CLEF [8] is a shared task in the field of cyber forensics that started in 2013, which does such analysis. Social media data like Tweets, comments etc, when analyzed, gives information about such fake accounts. Such research has been carried out for data in many languages, but analysis of Indian languages has only made its baby steps till now. One such step in the analysis of Indian languages, was the shared task of "Indian Native Language Identification (INLI)", held from 8th to 10th of December, 2017, at IISc Bangalore, India [1].

2 Related Work

The most relatable works are the systems submitted [8] for the PAN@CLEF 2017 Author Profiling shared task. Of the 20 teams that submitted their notebooks, the system that secured the first place [2] has used a linear SVM as the classifier. The aspects used for classification are word uni-grams along with character 3 to 5-grams [11]. Other features include POS tags and Twitter handles, along with geographic entities, but these proved to be inefficient in increasing the accuracy. The Twitter 14k and PAN@CLEF 2016 datasets were also used to improve the accuracy, but in vain. After trying different tokenization techniques, they settled for the scikit learn tokenizer, since the former did not improve the average accuracy. Apart from this, they had also analyzed emojis and did POS tagging on the data, only to get lower accuracy than the simple initial model.

The system which secured the second place [9] performed the following data preprocessing - non sense Tweet removal where English words' spelling was checked and wrongly spelled words were removed, and reversal of Arabic data. Preprocessing specific to Twitter data include removal of stop words, punctuation, hashtags etc. Of the classifiers used, Logistic Regression gave the highest accuracy. Other classi-

fiers used were Random forest and XGBoost. Linear SVM has also been used, but underperformed, when compared to Logistic Regression. A combination of classifiers like Logistic regression combined with Voting classifier was also tried but eventually, the best results were obtained with Logistic regression.

The system which secured the third place in the PAN@CLEF 2017 shared task [7] uses a MicroTC, which is a generic framework for text classification task, i.e., it works regardless of both domain and language particularities. They have used binary and trivalent parameters for preprocessing. They have used tokenizers and TFIDF weighting, all to be finally classified using a simple linear SVM. Hence, of the top 3 papers submitted for PAN@CLEF 2017, classic machine learning algorithms have scored the best. Of the other notebooks submitted, Barathi Ganesh HB et al. [3] uses Vector Space Models for text classification. Bougiatiotis K et al. [6], submitted for the PAN@CLEF 2016, uses stylometric features for the classification.

3 Dataset

The data set consists of Twitter data, in XML format. The corpora is the training data set of the shared task of Author Profiling, at PAN@CLEF 2017 [8]. The gender and language varieties selected are shown in Table 1. For each variety, the capital of the region where this variety has been used is selected. After selecting the region, Tweets in a radius of 10km from this region has been collected. PAN, being a team focused on research in this specific area, has overcome challenges like authenticity of authors, because of the previously collected data, where the uniqueness of each author is ensured. The time line of each author is retrieved which provides information such as the author's official name, her/his language and also the location.

Table 1 Gender and Language Variety of PAN@CLEF 2017 corpora

Language	Variety	Gender
English	Australia, Canada, Great Britain, Ireland, New Zealand, United States	Male, Female
Spanish	Argentina, Chile, Colombia, Mexico, Peru, Spain, Venezuela	Male, Female
Portuguese	Brazil, Portugal	Male, Female
Arabic	Egypt, Gulf, Levantine, Maghrebi	Male, Female

For each author, it is ensured that at least 100 tweets are there, with no re-Tweets. It is also ensured that the Tweets are written in the required variety of language. An author is concluded to be using a particular language variety if the Tweets' location is where this language is prevalent. Using a dictionary consisting of proper nouns,

automatic assigning of gender is done for an author. Otherwise, it has been done manually. A total of 500 authors' tweets have been collected, for language variety, as well as for gender. The training data set released, consists of 60% of the corpus formed, for language variety and gender.

Each author's XML file consists of 100 tweets. For each language, the corresponding truth labels are also provided, for both gender and language variety. The no. of XML files in English, Spanish, Portuguese and Arabic are 3600, 4200, 1200 and 2400 respectively. Hence, there are 3600x100 English Tweets, 4200x100 Spanish Tweets, 1200x100 Portuguese Tweets and 2400x100 Arabic Tweets.

4 Methodology

The basic methodology used for gender identification and language variety identification is as given in Figure 1, implemented using scikit-learn. The corpora given is in XML format, which is first converted to text format with the help of Document Object Model of XML data. It is then converted to matrix format, either as a DTM or as a TF-IDF matrix. The features of this text are further learned using minimum document frequency, word n-grams and character n-grams. Using the parameter "sub linear term frequency weighted TFIDF matrix" [2], where instead of normal term frequency, 1 + log(term frequency) is taken, has not improved the accuracy much. Then the classification is done using Machine Learning algorithms like SVM, AdaBoost, Decision Tree and Random Forest. Since the test data for PAN@CLEF 2017 Author Profiling has not been released yet, a 10 fold cross validation has been done on the training data and the accuracy is recorded.

4.1 Text Representation and Feature Learning

The first step in NLP, is to represent text in a format on which Machine Learning techniques can be easily applied. The formats in which text is represented in various ways are Distributed representation, Distribution Representation [4], Vector Space Models [3], [12]. Here, the count based methods of the Distribution Representation are used:

4.1.1 Term Frequency - Inverse Document Frequency (TF-IDF)

TF-IDF matrices are used when frequently used words need not be considered for the final analysis of text [12]. The rows represent documents and the columns represent vocabulary. It gives importance to unique words by transforming the mere count of a word, to the probability of the occurrence of the word in each document, by dividing it with the total no. of occurrences in all documents. It can be mathe-

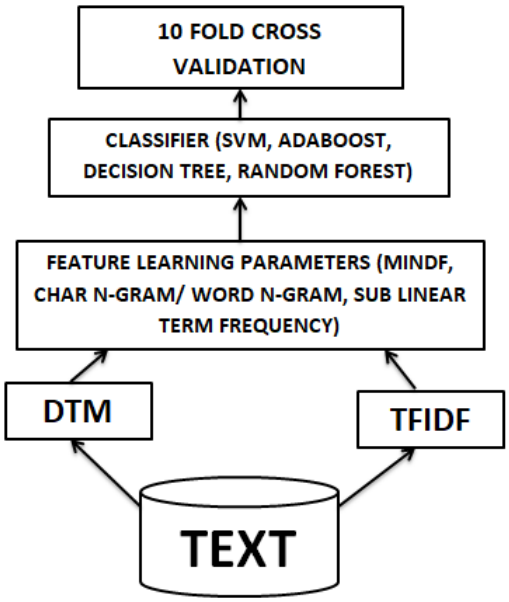

Fig. 1 ARCHITECTURE

matically expressed as:

$$TF - IDF = log\frac{N}{n_i} \tag{1}$$

where N is the no: of documents under consideration and
n_i is the no: of times the i^{th} word is occurring in the document.

4.1.2 Document Term Matrix (DTM)

DTMs directly take the count of each word occurring in each document and enters it into the matrix [12]. The rows represent documents and the columns represent the vocabulary. When a document is encountered, the unique words are entered into the vocabulary and the count is entered into the matrix. The same procedure is followed for all documents, hence increasing the vocabulary with each document encountered. It can be expressed mathematically as:

$$DTM = \frac{f_{t,d}}{\sum_{t'd} f_{t',d}} \tag{2}$$

where $f_{t,d}$ is the frequency of the term 't' in document 'd'.

Author Profiling is a task where the no. of words used by authors helps in knowing their gender, as most females use more words to express themselves. Even the information about the usage of stop words can end up being beneficial in Author Profiling. Hence, DTMs are more useful than TFIDF matrices, for this task.

4.1.3 Parameters for Feature Extraction

The parameters used for feature learning are as given below. Their contribution is crucial for the prediction [10].

1. Minimum Document Frequency (*mindf*) - It is specified to set a threshold to avoid those words which do not occur in at least "n" documents, where mindf = n.
2. N-gram - It gives the continuous occurrence of 'n' characters or words in a text, hence classified as character n-grams and word n-grams respectively.
3. Analyzer - N-grams can be applied for characters as well as for words, which is specified using the "analyzer" parameter. Character n-grams are mostly useful when the language dealt with is unknown.
4. Sub-linear Term Frequency - For TFIDF matrices, a parameter called sub linear term frequency can be specified, where instead of taking the term frequency directly, 1 + log(term frequency) is taken.

Table 2 shows the values taken up by different parameters

Table 2 Feature Extraction Parameters

Parameter	Values
mindf	1 - 25
analyzer	char, word
sublinear_tf	True, False
n-gram	1,2,3,4,5

4.2 Classifiers for Author Profiling

Machine Learning is a field in Computer Science which trains computers to learn like humans. It can be applied to different forms of data - speech signals, images, stock market data, weather data and yes, text data also. Even if Deep Learning, which is a part of Machine Learning, has proved to be more efficient than traditional Machine Learning algorithms, for image data, stock market data etc, text data analysis still has an affinity for traditional Machine Learning algorithms. The same affinity is observed for the analysis of PAN@CLEF 2017 train data from the top ranking systems [8].

4.2.1 Support Vector Machine

SVM [14] is a supervised learning algorithm, where the machine is trained using a set of training data along with its labels, and tested using test data. The accuracy is tested by comparing these true labels and predicted labels. SVMs construct hyperplanes to separate the classes, such that there should be maximum separation between classes. This is done so that when new data is encountered for classification, the chances for making error in the prediction is kept minimal. The hyperplane has 2 bounding planes which lie on its either sides, and the points falling on these bounding planes are called support vectors.

4.2.2 Decision Tree

Decision Trees [14] are supervised Machine Learning algorithms, used mainly for classification of data. It has a structure similar to flow charts, where the top node is the root node and each internal node denotes a test on a feature of the data. The branches represent the outcome of a test and each leaf node holds a class label. It is simple to visualize and understand, and can handle all kinds of data. Non linear relations will not affect the persona of the tree. But the algorithm may be biased and may become unstable since a small change in the data will change the structure of the tree.

4.2.3 Random Forest

Random Forest [14] is one of the most popular and efficient Machine Learning algorithms, used for regression as well as classification. As the name suggests, it contains a combination of Decision Trees. As the number of Decision Trees increase, the prediction becomes more robust and hence more accurate. Each Decision Tree gives a class to which a data belongs to. The class to which the most number of Decision Trees classify the data to be in, is assigned as that particular data's class, and accuracy is checked. It is capable of handling datsets of higher dimensionality. An important feature is that it handles missing values, and maintains the accuracy also. Yet, it is like a black box in the matter of having control over what the model does.

4.2.4 Adaboost

In Machine Learning, bagging algorithms are known to increase the veracity of Machine Learning algorithms [14]. Boosting is necessarily a simple variation of bagging algorithms. It improves the learners as it emphasizes on areas where the classifier is not performing as expected. It involves a repetitive process where the training data is randomly split, to form a bag of data. This bag of data is used for

training, and the corresponding model is tested using the entire training data. While forming the next bag of data, the previous wrongly classified data are deliberately included in it and the same process is repeated as above. These steps when repeated, improves the overall accuracy.

Finally, 10 fold cross validation is done on the data, where the data is divided in the ratio of 9:1 for train:test, and the accuracy is recorded and analyzed.

5 Experiments and Results

Here, the results obtained, for gender identification as well as for language variety identification has been presented. 10 fold cross validation has been used for all algorithms. Character n-grams have been used where word n-grams may not be efficient, mostly in unknown languages.

5.1 Gender

The results of the the task of gender identification is as given in Table 3 and Table 4. Table 3 describes the results for Tweets of all the four languages, when the text is represented as DTM. Table 4 gives the results for the same Tweets, when the text is represented TF-IDF matrix. It can be observed that AdaBoost and Random Forest classifiers stand out in performance.

Table 3 Gender Identification: DTM

Classifier / Language	SVM	AdaBoost	Decision Tree	Random Forest
English	77.77	76	65.75	74.83
Spanish	72.50	73.16	63.26	69.35
Portuguese	78.25	79.08	71.667	73.916
Arabic	69.708	71.66	66.916	69.79

Table 4 Gender Identification: TFIDF

Classifier / Language	SVM	AdaBoost	Decision Tree	Random Forest
English	48.44	76.44	64.65	73.11
Spanish	48.52	73.19	63.45	67.73
Portuguese	46.08	80.75	69.66	71.916
Arabic	51	74	66.04	70.58

5.2 Language Variety

The results of the the task of language variety identification is as given in Table 5 and Table 6. Table 5 describes the results for Tweets of all the four languages, when the text is represented as DTM. Table 6 gives the results for the same Tweets, when the text is represented TF-IDF matrix. It can be observed that AdaBoost and Random Forest classifiers stand out in performance.

Table 5 Language Variety Identification: DTM

Classifier / Language	SVM	AdaBoost	Decision Tree	Random Forest
English	68.61	77.19	73.916	76.08
Spanish	82.38	84.595	84.095	83.904
Portuguese	98.16	98.66	96.75	98.50
Arabic	73.20	73.58	71.708	76.62

Table 6 Language Variety Identification: TFIDF

Classifier / Language	SVM	AdaBoost	Decision Tree	Random Forest
English	14.25	75.83	72.88	77.83
Spanish	12.619	83.47	81.976	84.09
Portuguese	50.916	98.916	96.58	98.33
Arabic	24.916	72.58	71.54	75.16

Hence, it can be concluded from these results that although SVM is a good classifier, AdaBoost and Random Forest classifiers outperform SVM in most cases pertaining to this dataset.

6 Conclusion and Future Work

The corpora used here, is the PAN@CLEF 2017 corpora for Author Profiling. It consists of Tweets, in four languages, namely, English, Spanish, Portuguese and Arabic, in the form of XML files. To identify gender and language variety, the data has been represented as DTM as well TF-IDF matrix. Feature learning has been done with parameters like *mindf*, word n-gram and character n-grams. Classic Machine Learning techniques such as SVM, AdaBoost, Decision Tree and Random Forest have been used for the classification purposes. It has been observed from the results that when using SVM as the classifier, DTM gives more accuracy than TF-

IDF, for both classifications. Also, the other classifiers - AdaBoost in particular - outperform SVM for this data set.

Another area of research is to analyze code mixed data, obtained from social media platforms like Facebook, Twitter and Whatsapp. Code mixed data is the combination of two languages, mostly English combined with other native languages, and is the common language of communication in social media. An analysis of the Indian code mixed language, Malayalam-English code mixed data to be specific, is of high prospect, as a future work. Deep Learning and Fast Text can be used for text classification [13].

References

1. Anand Kumar, M., Barathi Ganesh, H.B., Singh, S., Soman, K.P., Rosso, P. Overview of the INLI PAN at FIRE-2017 track on Indian native language identification (2017) CEUR Workshop Proceedings, 2036, pp. 99-105.
2. Angelo Basile, Gareth Dwyer, Maria Medvedeva, Josine Rawee, Hessel Haagsma, and Malvina Nissim, 12 Jul 2017, "N-GrAM: New Groningen Author-profiling Model, Notebook for PAN at CLEF 2017", CLEF 2017.
3. Barathi Ganesh HB, Anand Kumar M, Soman KP, 2017, " Vector Space Model as Cognitive Space for Text Classification", Notebook for PAN at CLEF 2017, CLEF 2017.
4. Barathi Ganesh, H.B., Anand Kumar, M., Soman, K.P. Distributional semantic representation for text classification and information retrieval (2016) CEUR Workshop Proceedings, 1737, pp. 126-130.
5. Barathi Ganesh, H.B., Reshma, U., Anand Kumar, M., Soman, K.P. Representation of target classes for text classification - AMRITA-CEN-NLP@RusProfiling PAN 2017 (2017) CEUR Workshop Proceedings, 2036, pp. 25-27.
6. Bougiatiotis, K., Krithara, A., "Author profiling using complementary second order attributes and stylometric features. CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers", CLEF 2016.
7. Eric S. Tellez, Sabino Miranda-Jimnez, Mario Grafi, and Daniela Moctezuma, 2017, "Gender and language-variety identification with MicroTC, Notebook for PAN at CLEF 2017", CLEF 2017.
8. Francisco Rangel, Paolo Rosso, Martin Potthast, Benno Stein, 2017, "Overview of the 5th Author Profiling Task at PAN 2017:Gender and Language Variety Identification in Twitter", CLEF 2017.
9. Matej Martinc, Iza krjanec, Katja Zupan, Senja Pollak, 2017, "PAN 2017: Author Profiling - Gender and Language Variety Prediction, Notebook for PAN at CLEF 2017", CLEF 2017.
10. Medvedeva, M., Kroon, M., Plank, B., "When sparse traditional models outperform dense neural networks: the curious case of discriminating between similar languages." In: Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects. pp. 156–163. Association for Computational Linguistics (2017).
11. Rangel, F., Franco-Salvador, M., Rosso, P.: "A low dimensionality representation for language variety identification.", CICLing - Computational Linguistics and Intelligent Text Processing, 2016.
12. Turney, P. D., Pantel P, "From Frequency to Meaning: Vector Space Models of Semantics", Journal of Artificial Intelligence Research, 2010.
13. https://www.analyticsvidhya.com/blog/2017/07/word-representations-text-classification-using-fasttext-nlp-facebook/
14. https://in.udacity.com/