Tlemcen University: Bots and Gender Profiling Task
Notebook for PAN at CLEF 2019

Rabia Bounaama[1] and Mohammed El Amine Abderrahim[2]

[1] Biomedical Engineering Laboratory, Tlemcen University,Algeria
rabea.bounaama@univ-tlemcen.dz
[2] Laboratory of Arabic Natural Language Processing, Tlemcen University, Algeria
mohammedelamine.abderrahim@univ-tlemcen.dz

Abstract This is about the participation of techno team at PAN @ CLEF 2019. We use to solve the task text analysis techniques and machine learning approaches. We describe the properties of our multilingual system based on Stochastic Gradient Descent (SGD) learning classifier submitted for PAN2019, which recognizes bots and gender profiling using tweets in two languages, namely, English and Spanish. We show the useful of some features to identify the text style and author's information. And then we evaluate the model on a number of unseen data sets. The proposed models have as accuracies 0.50, 0.25 for English prediction of a bots or human as well gender respectively.

Keywords: bots and gender profiling, machine learning, SGD classifier.

1 Introduction

Social media become one of the most popular ways for people to communicate and to post. Posts are generally variable in length and may involve multiple topics. An author's writing style can be affected by different topics and different replies/comments (e.g. supportive, negative and aggressive) [8]. In marketing, companies and resellers would like to know the view point of people about their products based on the analysis of blogs and product reviews [10], also people tend to seek out and receive news from it so these communications and ratings can produce significant quantities of data which must be analyzed.

These media allow hiding the real profile of the users who interact and generate information. Therefore, the possibility of knowing social media users traits on the basis of what they share is a field of growing interest named author profiling [11]. Author profiling deals with deciphering information about the author from the text that he/she has written [1], this helps in identifying aspects about the user.

Bots could artificially inflate the popularity of a product by promoting it and/or writing positive ratings, as well as undermine the reputation of competitive products through negative valuations[3]. Bots and Gender Profiling task at PAN 2019 CLEF [3,2]

[3] https://pan.webis.de/clef19/pan19-web/author-profiling.html

aim to determine whether the author of a tweet is a bot or a human. In case of human, identify her/his gender, the task is held in English and Spanish. Thus, the participants must provide their multi-lingual model solution to the problem. The performance of participants systems will be ranked by accuracy through TIRA [9].

The paper is structured as follows. In the next section, we give a brief overview of some related work. Section 3 describes the methodology and corpus preprocessing. Section 4 presents the results. Then we conclude the paper.

2 Related Work

Some of the recent studies in social media [1]where the authors propose a multi-lingual model for indentification of age and gender at PAN 2015 as classification task whether they apply a linear model SGD learning, and another Multilingual Personality prediction model where they apply a multivariate regression model of Ensemble of Regressor Chains Corrected (ERCC). Besides that in another work of author profile at PAN 2016 [4] where they used SVM-based classifiers, liblinear for gender classification and libsvm with a radial basis kernel to predict the age. Also they use NRC Word-Emotion Association Lexicon for their training data.

In [10] authors apply TF-IDF and a Deep-Learning model based on Convolutional Neural Networks. They compute the cosinus similarity between the T$\boldsymbol{f}$I$\boldsymbol{df}$ $\boldsymbol{d}$ vector and the vector T$\boldsymbol{fq}$ of term frequencies for their training data in order to predict the gender or language variety at PAN 2017 from unseen data test . Moreover in the work of [6] for the prediction of gender and language variety at PAN 2017 also in the work of [12]for the task of the past year (PAN 2018) concerns gender identification on Twitter we found that the authors use a logistic regression with good accuracy.

The studies mentioned above show the applicability of some statistical methods for author profiling tasks at PAN CLEF. In this paper we propose a multilinguale model for indentification of bots and gender profiling based on Stochastic Gradient Descent (SGD) learning classifier.

3 Method

In this section, we describe two multilingual predictive models that we use in our submission. We build a multilingual model for identifying bots or human users and a multilingual model for predicting their gender in case of human.

The organizers of PAN 2019 bots and gender Profiling Task provide a dataset which consisted of two different training sets for the different languages: English and Spanish for the total 412000,300000tweets respectively , collections is depicted in table1.

The data was given in the form of xml files containing tweets for several users. We apply the following set of preprocessing steps to all documents.

First we created a function to extract tweets from xml files and save them to a csv file using the "beautifullsoup"[4], "Pandas"[5] libraries for both languages. We used only

[4] https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[5] https://pandas.pydata.org/

Table 1: Training corpora statistics

language	Tweets	Authors (human / bots)		Gender (male / female)	
English	412000	206000	206000	103000	103000
Spanish	300000	150000	150000	75000	75000

the posts text for training containing the tweet only with the author and the author's gender we extract all tweets belonging to one user .We performe a pre-processing for the data before used it to train SGD(svm) classifier. The following pre-processing were performed:

- Removing url links, @ username,Hashtag# .
- Tokenizing text by white space.
- Normalizing case to lowercase.
- Removing punctuation from each word.
- Removing non-printable characters.
- Removing stopwords.
- Lemmatizing words .

Secondly, we have started the stage of the construction of the model, in this stage we have created three functions first of all the creation of the classifier from which it takes as parameter the specified classifier, the vector of features of learning with the output classes and the validation vector.

According to [5], the use of N-grams is the best method to analyze emotions in microblogging context. So we train our classifier by using 3-grams features. From these features, we selected only those that have as minimum term count frenquency equal to 3 in the classification task and we used them in the third function in order to train the model.

We used the same presentation of features and model parameters as the ones chosen for English to train Spanish dataset. Our model was built using the tools provided by scikit-learn machine learning library in Python [7]. We also tested several classifiers and different parameter sets. The following classifiers from Scikit-learn were tested:

- Svm.linearSVC
- Logistic regression
- RNN (reccurent nereunal network)
- Naïve bayes multinominal

Best results were obtained with SGD classifier, we used 'hinge' as loss function and L2 for penalitie, to our submitted run .

4 Results

For the task of bots and gender profile prediction, we obtain better results for the prediction of Spanish language as presented at table 2 and 3.

Table 2: Gender prediction

language	English	Spanish
Accuracy	0.2511	0.2567

Table 3: Bots/human prediction

language	English	Spanish
Accuracy	0.5008	0.5050

Our techno team have as an average of score 0.3784 . According to the obtained results we found that sgd (svm) classifier perform better for author prediction while this approach did not perform very well at gender prediction.To overcome this limitation, we plan to do more advanced preprocessing using, for example, linguistic markers.

We faced some limitation in building our system such as :

- Tweets data contains incorrectly words for example people spell the word "soon" as "soooon" to convey excitement in such situations, tokenizing and identifying words becomes challenging.
- Social media users use their own vocabulary to express their thoughts or feeling, thus extracting vocabulary-based or grammar-based features may not work efficiently for these platforms. Furthermore, social media users use multiple languages to express their opinion. This makes it impossible to apply knowledge derived from one language by extracting language dependent features, onto another language.

5 Conclusion

We have presented the system developed by our techno team for participating in PAN-2019 bots and gender profiling Task, we designed and implemented a system that could be easily configured where we use in our final model SGD classifier. The main challenge with this model is then to fight effectively overfitting. The biggest challenge of this year's PAN bots and gender profiling task was the gender classification problem where our model achieves an average of 0.25 accuracy.

References

1. Mounica Arroju, Aftab Hassan, and Golnoosh Farnadi. Age, gender and personality recognition using tweets in a multilingual setting. In *6th Conference and Labs of the Evaluation Forum (CLEF 2015): Experimental IR meets multilinguality, multimodality, and interaction*, pages 22–31, 2015.
2. Franco M. Francisco Rangel, Paolo Rosso. A low dimensionality representation for language variety identification. In *Proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'16)*. Springer-Verlag,LNCS(9624),pp. 156-169, 2018.
3. Paolo Rosso Francisco Rangel. Overview of the 7th author profiling task at pan 2019: Bots and gender profiling. CEUR Workshop Proceedings, In: Cappellato L., Ferro N., Müller H, Losada D. (Eds.), 2019. CEUR-WS.org <http://ceur-ws.org>.
4. Pepa Gencheva, Martin Boyanov, Elena Deneva, Preslav Nakov, Yasen Kiprov, Ivan Koychev, and Georgi Georgiev. Pancakes team: A composite system of genre-agnostic features for author profiling. In *CEUR Workshop Proceedings*, 2016.

5. Gonzalo Blázquez Gil, Antonio Berlanga de Jesús, and José M. Molina Lopéz. Combining machine learning techniques and natural language processing to infer emotions using spanish twitter corpus. In *Highlights on Practical Applications of Agents and Multi-Agent Systems*, pages 149–157. Springer Berlin Heidelberg, 2013.
6. Matej Martinc, Iza Skrjanec, Katja Zupan, and Senja Pollak. Pan 2017: Author profiling-gender and language variety prediction. In *CLEF (Working Notes)*, 2017.
7. Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python journal of machine learning research. 2011.
8. Jian Peng, Kim-Kwang Raymond Choo, and Helen Ashman. Bit-level n-gram based forensic authorship analysis on social media: Identifying individuals from linguistic profiles. *Journal of Network and Computer Applications*, 70:171–182, 2016.
9. Martin Potthast, Tim Gollub, Matti Wiegmann, and Benno Stein. TIRA Integrated Research Architecture. In Nicola Ferro and Carol Peters, editors, *Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF*. Springer, 2019.
10. Nils Schaetti. Unine at clef 2017: Tf-idf and deep-learning for author profiling. In *CLEF (Working Notes)*, 2017.
11. Mariona Taulé, M Antonia Martí, Francisco M Rangel, Paolo Rosso, Cristina Bosco, Viviana Patti, et al. Overview of the task on stance and gender detection in tweets on catalan independence at ibereval 2017. In *2nd Workshop on Evaluation of Human Language Technologies for Iberian Languages, IberEval 2017*, volume 1881, pages 157–177. CEUR-WS, 2017.
12. P von Daniken, Ralf Grubenmann, and Mark Cieliebak. Word unigram weighing for author profiling at pan 2018. In *Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*, 2018.