Bots and Gender Profiling using Character and Word N-Grams

Notebook for PAN at CLEF 2019

Mahendrakar Srinivasarao and Siddharth Manu

msrini.svn@gmail.com
siddharthmanu94@gmail.com

Abstract Author profiling, a term used for analysing of text and identifying characteristics of a person based on stylistic and content-based features. In this paper, we describe the approach to detect bot and human (male or female) out of the authors of tweets as a submission for Bots and Gender Profiling shared task at PAN 2019. Our approach involves a combination of character and word n-grams as features for each class and trained Support Vector Machine (SVM). Our experiments show that this method gives good performance in detecting bot and gender (male or female).

1 Introduction

Bots played a key role in generating large amounts of internet traffic in the recent years, in fact they have become ubiquitos in the social media platforms like Twitter, Facebook, etc [15]. Social media bots pose as human to influence users with commercial, political or ideological purposes. For example, 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 . The threat is even greater when the purpose is political or ideological [1]. Research shows that in 2016 U.S. Presidential Election, more than 1/5 of tweets on Twitter came from bot accounts [4]. Furthermore, bots are commonly related to fake news spreading [7]. Therefore, bot detection on social media, especially on Twitter has become an important research area across the globe. This year's shared task on bots and gender profiling at PAN 2019 [12], aims to investigate whether the author of a Twitter feed is a bot or a human. Furthermore, in case of human, to profile the gender of the author in two different languages English and Spanish.

Bot and gender classification is binary problem and with in the gender, male or female classification is again a binary classification. In this paper, we present our approach in the final submitted software version at TIRA platform [2].

2 Related Work

Word and character n-grams have been strong predictors of gender in author profiling[9]. For author profiling, it has been shown that tf-idf weighted n-gram features, both in terms of characters and words, are very successful in capturing especially gender distinctions [14], [6]. Character and word grams have proven to obtain decent results in gender classification on Twitter. In the paper [5] authors use word unigrams, bi-grams and character 1-5 grams as features to feed into various training algorithms. Most of the best performing teams in author profiling task at PAN have adopted similar approaches to obtain good accuracies [3], [6]. In the past years shared tasks at PAN, traditional machine learning training algorithm Support Vector Machine (SVM) has been used in combinations of character and tf-idf word n-grams [13]. Even though there are two different tasks here(one bot/gender and other male/female), can a model be built with the same set of features that are used extensively for gender detection for bot/gender detection as well ?

3 Dataset and Preprocessing

The dataset provided consists of a series tweets in the form of XML files, each one corresponding to an author, containing 100 tweets. Tweet text is in raw format, containing links, mentions to other users and hashtags.
Two groups of dataset are provided.
English: 4,120 authors,
Spanish: 4,120 authors
Each XML file per author (Twitter user) with 100 tweets and authors were coded with an alpha-numeric author-ID.

Most of the preprocessing is done with the of *TweetTokenizer* module of the Natural Language Took Kit library. Approaches followed in preprocessing tweet text are similar to commonly used techniques [8] and [6].

- Replacing line feed with <*LineFeed*>
- Tweet concatenation into one for a single author
- Replace URL with <*URLURL*>
- Removal of punctuations
- Trim repeated character sequences of length >= 3

4 Features

In author profiling task, PAN 2018, second best performing team [6] used different combinations of word and character n-grams on tweet text. This has motivated us to use the similar approach for the bot and gender detection task as well. Table 3 shows character and word n-gram hyper parameters used which are obtained after different experiments on both English and Spanish datasets.

TF-IDF matrix created out of character and word n-grams (term frequencey of less than 2 omitted). Dimensionality reduction on this matrix is done using Singular Value

Table 1. n-Gram Hyper Parameters used for English and Spanish

Language/n-grams	**English**	**Spanish**
Character grams	*3-4*	*3-4*
Word grams	*1-3*	*1-2*

Decomposition (SVD) and library call truncateSVD from scikit learn was used. The reduced rank space contained only 200 features as optimal. Increasing in number of componets (> 200) in the reduced rank space resulted in decreased accuracy and sometimes resulted in memory error on 4GB RAM Tira virtual machine. Support Vector Machines (SVM) has been proven to obtain decent results in author profiling [6], [9] . When compared with other trainers SVMs proved to be more discriminatory. Therefore, the implementation of linear SVM in the library scikit-python [10] was chosen as the classification method. In order to prevent overfitting, the value of C was fixed in 1.0, as done in [15].

4.1 Experiments and Results

In order to validate the approach, the data for each language was split in 60% for training and 40% for test (i.e 2472 documents for training and 1648 for testing). The experiments are made from a subset, the classification in the final task will be made using all the training data. We have tried different trainers NaiveBayesPredict, LogisticRegression and LinearSVC. Model training is done using 10-fold cross validation as it has obtained good results [6]. LinearSVC is choosen in the final version of the software as it has given good results over the others. Results on test data (which is 40% of the original training data) are shown in Table 2 for English dataset. In the final submission, model is trained

Trainer Used	CrossValidataion Mean Accuracy	TestSet Accuracy
NaiveBayesPredict	66.69	58.37
Logistic Regression	92.39	90.23
LinearSVC	94.42	93.08

Table 2. Accuracy on English Test-set (40% of training data).

on the whole training set using SVM Classifier and tested on the official PAN 2019 test set for the author profiling task, on the TIRA platform [11]. Results obtained on final submission are shown in Table 3.

Table 3. Results obtained on Final Test Data Set

Language	BOTS vs. HUMAN	Gender	Average
English	0.9371	0.8398	0.8884
Spanish	0.9061	0.7967	0.8514
Average	0.9216	0.8182	**0.8699**

5 Conclusion

The simple approach defined here and in the past [6] performs well when compared with others, decently. Word unigram and bigrams have given good results and increasing word n-gram size beyond 2 decreased the performance for both English and Spanish datasets. This hyper parameter tuning was necessary. Initial submission of software resulted in memory error due to more number of components in reduced rank space (done using truncatedSVD). However, increasing the number of components beyond 200 did not improve the performance. SVM still remains at the top for the bot/gender detection task based on our experiments. As a future work, deep neural networks can be considered, especially Convolutional Neural Networks (CNN) to obtain better results.

References

1. Bots and gender detection (2019), https://pan.webis.de/clef19/pan19-web/author-profiling.html
2. Tira platform (2019), https://www.tira.io/
3. Basile, A., Dwyer, G., Medvedeva, M., Rawee, J., Haagsma, H., Nissim, M.: N-gram: New groningen author-profiling model. arXiv preprint arXiv:1707.03764 (2017)
4. Bessi, A., Ferrara, E.: Social bots distort the 2016 us presidential election online discussion (2016)
5. Burger, J.D., Henderson, J., Kim, G., Zarrella, G.: Discriminating gender on twitter. In: Proceedings of the conference on empirical methods in natural language processing. pp. 1301–1309. Association for Computational Linguistics (2011)
6. Daneshvar, S., Inkpen, D.: Gender Identification in Twitter using N-grams and LSA: Notebook for PAN at CLEF 2018. In: CEUR Workshop Proceedings. vol. 2125 (2018), http://ceur-ws.org/Vol-2125/paper_213.pdf
7. Fox, M.: Want something to go viral? make it fake news. https://www.nbcnews.com/health/health-news/fake-news-lies-spread-faster-social-media-truthdoes-n854896 (2018)
8. Magliani, F., Fontanini, T., Fornacciari, P., Manicardi, S., Iotti, E.: A comparison between preprocessing techniques for sentiment analysis in twitter (12 2016)
9. Oliveira, R.R., Neto, R.F.O.: Using character n-grams and style features for gender and language variety classification. In: CLEF (Working Notes) (2017)
10. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. Journal of machine learning research 12(Oct), 2825–2830 (2011)
11. Potthast, M., Gollub, T., Wiegmann, M., Stein, B.: TIRA Integrated Research Architecture. In: Ferro, N., Peters, C. (eds.) Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF. Springer (2019)

12. Rangel, F., Rosso, P.: Overview of the 7th Author Profiling Task at PAN 2019: Bots and Gender Profiling. In: Cappellato, L., Ferro, N., Losada, D., Müller, H. (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2019)
13. Rangel, F., Rosso, P., Potthast, M., Stein, B.: Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. Working Notes Papers of the CLEF (2017)
14. Sanchez-Perez, M.A., Markov, I., Gómez-Adorno, H., Sidorov, G.: Comparison of character n-grams and lexical features on author, gender, and language variety identification on the same spanish news corpus. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 145–151. Springer (2017)
15. Zampieri, M., Tan, L., Ljubešić, N., Tiedemann, J., Nakov, P.: Overview of the dsl shared task 2015. In: Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects. pp. 1–9 (2015)