Bots and Gender Profiling Using a Deep Learning Approach

Notebook for PAN at CLEF 2019

Jose R. Prieto Fontcuberta and Gretel Liz De la Peña Sarracén*

Universitat Politècnica de València
{joprfon,gredela}@posgrado.upv.es

Abstract This paper describes the system we developed for the Bots and gender profiling task, at PAN @ CLEF 2019. The task consists in, given a tweets set, automatically determine whether its author is a bot or a human. In case of human, identify her/his gender. We propose a deep learning based system, fed with the TFIDF representation from the texts instead of word embeddings representation as usual. Additionally, we use some linguistic features which improve the performance of the system according with the experimental results.

1 Introduction

Nowadays we use a lot of social media content, being a powerful tool to communicate to the world. Some enterprises use bots accounts as a social manager to answer fast, free and automatically to their clients. However, sometimes some governments, people or powerful institutions abuse of these social networks and create bots to manipulate and distortion the information and the point of view of some users [3], [2]. A bot can be defined as a program that learns to promote some information as a normal user but automatically, and can be programmed with a software specially concerned on the manipulation on some topics. Hence identifying bots in the social networks is a relevant task, not only from a point of view of marketing, but also from a forensics and security perspective.

Among the efforts made to address this phenomenon, this year the Bots and gender profiling task [9] has been launched as part of PAN @ CLEF 2019[1] [1]. This task focuses on detecting bots against humans users given a text set from Twitter, one of the most used social networks. The tweets are in English and Spanish.

In recent related works as in [4], deep learning techniques, used for text classification purposes, are used also for this task. More recent and new techniques are explained and used in [5], where also Word Embeddings, dense layers and LSTM are used. In general, many works point out the flexibility in capturing nonlinear relationships of deep learning techniques.

* The authors contributed equally to this paper.

[1] https://pan.webis.de/clef19/pan19-web/

In this paper, we propose a Feedforward Neural Network of two layers for the task. In addition, as a second step, the system should identify the author gender in case of human. For this other task we use a similar architecture but in this case with 4 layers. A point to highlight is the use of some linguistic features which can help to discriminate among types of users. The paper is organized as follows. Section 2 describes our system. Experimental results are then discussed in Section 3. Finally, we present our conclusions with a summary of our findings in Section 4.

2 System

2.1 Preprocessing

In the preprocessing step the text is cleaned. Firstly, the typical characteristics used in the tweets, and that possibly do not have discriminatory semantic information, have been identified. We identify urls, numbers, mentions to users and dates as that kind of features. Then, each part of the texts which represents this kind of features is replaced with a corresponding tag.

2.2 Method

We propose a model that consists in a Feedforward Neural Network (FFNN) with two layers, BatchNorm and ReLU activation. As input, the model takes a vector which is generated as TFIDF representation of the concatenation of all tweets of the user. Thus, we achieve to represent the information of a user in a unique vector. Given a text set S, a term t, and an individual text $T \in S$, we calculate:

$$TFIDF(t) = tf(t,T) * log(\frac{|S|}{tf(t,S)}) \quad (1)$$

Where $tf(t,T)$ is the number of times t appears in T, $|S|$ is the size of the corpus, and $tf(t,S)$ is the number of texts in which t appears in S.

This representation was selected based on the idea that the words usually used by a user and no commonly by others, could be more important in the corresponding representation of the user. This is an idea that matches with the phenomenon which we try to capture. That is, this kind of representation can allow us to get typical characteristics of each user.

We also tried to encode with TFIDF every tweet separately and then concatenate all of them, creating as a result something similar to a gray scale image. For this representation we used a 2D Convolutional Neural Network with large kernels matching with the number of features extracted from the TFIDF representation (2DCNN-TFIDF). These large kernels are so expensive even when a GPU is used, due to this the train is extremely slowly. We also tried to use a Recurrent Neural Network with a few layers of LSTM (LSTM-TFIDF). On both experiments the results were not enough satisfactory.

Similarly, we tried other approaches with other kinds of representation, using the GloVe Word-Embeddings. We use all the tweets concatenated and then we look up

into the Word-Embedding table. As before, for this representation we used a 2D Convolutional Neural Network with large kernels matching with the number of features extracted from the embeddings (2DCNN-WE) and a Recurrent Neural Network with a few layers of LSTM (LSTM-WE). Again, on both experiments the results were not enough satisfactory.

Finally, we tried to concatenate the tweets of a given user on depth, obtaining as result an *image* with the number of channels (depth) equivalent to the number of tweets of the user. Each matrix is the concatenation of words embeddings for a tweet. Therefore, the width and height are determined by the size of embeddings and number of words in the tweet, respectively. With this approach we used a 3D Convolutional Neural Network with larger kernels (3DCNN-WE). With this last approach the idea is to consider the kernels as large as the size of the embedding and in the second dimension we chose a n-gram value to consider the context.

2.3 Linguistic Features

We include some linguistic features which consider important to discriminate among users. For the implementation we used the TextBlob library[2] which can be used to process texts in English. Hence, these features were employed just with the corpus in English. Two types of features were analyzed. On one side, features related with subjective information and on the other hand, features related with syntactic information:

- Subjective information (SI): Analysis of degree of subjectivity and sentiment present in the text. This can be a good discriminative feature, since bots can be less subjective and sentimental than humans.
- Syntactic features (SF): Analysis of count of adjectives, adverbs and pronouns in texts. These kinds of features can discriminate between male and females as some studies suggest [6].

3 Results

In this section, we report and discuss the performance of the system in the task. Training and evaluation were conducted using the PAN @ 2019 proportioned datasets which have 4120 and 3000 tweets in English and Spanish respectively. The data is balanced for each subtask and language. Results are obtained by uploaded the system to *TIRA* [7].

The results obtained on the development set with each approach commented before are reported in Table 1. As we can see the best results are achieved with the FFNN method. Hence it was the system selected for the task at PAN 2019. Other models that supposed to be superior obtained worse results. Perhaps it could be due to the large number of introduced parameters that were not well trained due the small amount of data available.

[2] https://textblob.readthedocs.io/en/dev/

As we can see in Table 1, the best results achieved have an accuracy of 0.90 and 0.93 on Spanish and English datasets, respectively. In the Spanish partition we do not use linguistic features, as we do in English dataset, where we use SI and SF features.

Method	FFNN	2DCNN-TFIDF	LSTM-TFIDF	2DCNN-WE	LSTM-WE	3DCNN-WE
Accuracy ES	0.90	0.76	0.73	0.70	0.69	0.72
Accuracy EN	0.93	0.78	0.75	0.72	0.70	0.73

Table 1. Results of different approaches on the development dataset

In Table 2 we can see the improvement of the FFNN method adding the linguistic features (LF). As we commented before, we just used them for the English corpus. The system gains one point of accuracy adding these features, but there were no differences when adding the syntactic features.

Method	Without LF	With SI	With SI + SF
Accuracy EN	0.92	0.93	0.93

Table 2. Results in the Bot vs Humans Task with FFNN

Table 3 shows the results of accuracy for gender profiling for those tweets predicted as a human. We achieve 0.87 accuracy on the English corpus, and the results did not vary when the linguistic features are added. Hence these features are not important for this task according to our experimental results. On the Spanish corpus, 0.86 accuracy is achieved without any linguistic feature.

Method	Without LF	With SI	With SI + SF
Accuracy EN	0.87	0.87	0.87
Accuracy ES	0.86	-	-

Table 3. Results in the Gender Task with FFNN

Table 4 shows the results on the test datasets for our final model and the baselines proposed by the organizers of the task. We outperform in almost all the cases the results of the baselines, except in English bot task where the LDSE method [8] obtains better results.

Method		MAJORITY	RANDOM	LDSE	*Our proposal*
Bots	**EN**	0.5000	0.4905	0.9054	0.9045
	ES	0.5000	0.4861	0.8372	0.8578
Gender	**EN**	0.5000	0.3716	0.7800	0.7898
	ES	0.5000	0.3700	0.6900	0.6967

Table 4. Final results in both task on the test datasets

4 Conclusion

We propose a deep learning based system for the Bots and gender profiling task, at PAN @ CLEF 2019. The model consists of a Feedforward Neural Network which gets as input the TFIDF representation from the text. The experimental results show the suitability of the used representation for the task, achieving 0.8578 of accuracy on the Spanish corpus and 0.9045 on the English corpus, on detecting bots vs human. For gender profiling we obtain an accuracy of 0.6967 and 0.7898, respectively. Also, some linguistic features are added, allowing for a small improvement in the bots and human discrimination task, but not for gender profiling. Furthermore, we tried to use word embeddings and some different kinds of architectures with these new features but the results were not enough satisfactory. Maybe the results of CNN might be improved with more data for re-training the word-embedding for this specific task.

References

1. Daelemans, W., Kestemont, M., Manjavancas, E., Potthast, M., Rangel, F., Rosso, P., Specht, G., Stamatatos, E., Stein, B., Tschuggnall, M., Wiegmann, M., Zangerle, E.: Overview of PAN 2019: Author Profiling, Celebrity Profiling, Cross-domain Authorship Attribution and Style Change Detection. In: Crestani, F., Braschler, M., Savoy, J., Rauber, A., Müller, H., Losada, D., Heinatz, G., Cappellato, L., Ferro, N. (eds.) Proceedings of the Tenth International Conference of the CLEF Association (CLEF 2019). Springer (Sep 2019)
2. Ferrara, E.: Disinformation and Social Bot Operations in The Run Up to The 2017 French Presidential Election. arXiv preprint arXiv:1707.00086 22(8), 1–33 (2017)
3. Forelle, M., Howard, P., Monroy-Hernández, A., Savage, S.: Political Bots and The Manipulation of Public Opinion in Venezuela. CoRR abs/1507.07109 (2015), http://arxiv.org/abs/1507.07109
4. John, V.: A Survey of Neural Network Techniques for Feature Extraction from Text. arXiv preprint arXiv:1704.08531 (2017)
5. Kudugunta, S., Ferrara, E.: Deep Neural Networks for Bot Detection. Information Sciences 467, 312–322 (2018)
6. Nerbonne, J.: The Secret Life of Pronouns. What Our Words Say About Us. Literary and Linguistic Computing 29(1), 139–142 (2014)
7. 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)
8. Rangel, F., Franco-Salvador, M., Rosso, P.: A Low Dimensionality Representation for Language Variety Identification. In: International Conference on Intelligent Text Processing and Computational Linguistics. pp. 156–169. Springer (2016)

9. 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)