Stacked Bots and Gender Prediction from Twitter Feeds

Notebook for PAN at CLEF 2019

Todor Staykovski

Sofia University, Sofia, Bulgaria
todorstaykovski@gmail.com

Abstract This paper describes an approach of predicting bots and gender from twitter feeds in English only. Given an author Twitter feed, the task is to identify whether it is a bot or human and in case of human to profile the gender of the author. The submitted system is a stack of two models. In the first model, the feeds are represented as TF–IDF vectors. The second model is based on Doc2Vec feeds representations. Both models use the same pre-processing pipeline and multilayer perceptron neural network in order to classify the tweets.

1 Introduction

Over the last years bots have flooded social media, serving different purposes, such as commercial and political. In order to achieve their goal, bots are spreading fake news. Some bots promote specific products, while others could pose a thread for the security, spreading news or opinions on political or ideological bases. Thus identification of bots on social media is very important task nowadays.

In the past years PAN have organized several tasks on author profiling. This year's Author Profiling task at PAN 2019 [2] is different from the previous years, since classification is separated into two sub-tasks. The first sub-task is to identify whether the author of a Twitter feed is a bot or a human. The second sub-task is continuation of the first one, profiling the author in case of human. The performance of the submitted solutions is ranked bu accuracy. The accuracy is calculated for both sub-tasks separately, where the final accuracy result will be average per language. Further details about Bots and Gender Profiling task can be found in [8].

This paper presents an approach of predicting bots and gender, only for English language. The system predicts for both classification sub-tasks: (i) bot or human; (ii) bot, male or female.

The remaining of this paper is organized as follows: Section 2 gives a brief overview of relevant previous work. Section 3 describes the dataset. Section 4 presents the core system stack based on two models. Section 5 presents the experiments and discusses the evaluation results. Finally, Section 6 concludes and points to possible directions for future work.

2 Related Work

Author profiling has been explored in PAN lab for years. Wide range of machine learning algorithms have been presented, including deep learning.

Giovanni Ciccone *et. al.* [1] predict gender from tweet texts and images, using stack composed of two main parts. The first one, predicts gender from texts, based on n-grams and TF–IDF. The second part consists of different layers of classifiers, predicting gender from images. Both predictions, from texts and images, feed into a classifier that outputs the final prediction.

Adam Poulston *et. al.* [7] present an approach of predicting gender from tweets, based on stack of two classifiers: a logistic regression classifier trained on TF–IDF transformed n-grams and a gaussian process classifier trained on word embedding clusters.

3 Dataset

The English part of the training dataset of the Bots and Gender Profiling task consists of 4120 files. Every file contains 100 tweets from an unique author. The dataset is balanced for both sub-tasks. The labeling of the two sub-tasks is as follows:(i) classification by type: human or bot; (ii) classification by gender: bot, male or female.

The training dataset is further separated into train and development parts, which are balanced as well.

4 Overview of the Proposed System

This section describes the two parts of the stacked model. Both parts use the same preprocessing pipeline and multilayer perceptron neural network, where outputs one-hot vector. The output is bot, male or female which serves for both sub-tasks.

In order to prevent over-fitting on the training dataset a dropout is used. The value of the dropout $d = 0.2$ for both models of the stack was chosen empirically.

4.1 Pre-processing

The pre-processing step involves case-folding, parsing specific symbols to text and punctuation removing. All tweet specific symbols are parsed to text. For instance the sentence "Bots and Gender Prediction is #great competition :-) https://pan.webis.de" is pre-processed to "bots and gender prediction is hashtag competition emoji url". After cleaning the tweets they are tokenized using TweetTokenizer from the nltk library [4], using options of strip handles and reduce length.

4.2 Prediction Based on TF–IDF Vector Representations

The first part of the stack uses TF–IDF vector representations. In order to build the representations, a TF–IDF vectorizer from [5] is used. The vocabulary of the vectorizer is built from the training dataset. The TF–IDF vectorizer outputs a vector with size of 20,000. A feed-forward neural network with 2 hidden layers is trained, using hyperparameters shown in Table 1.

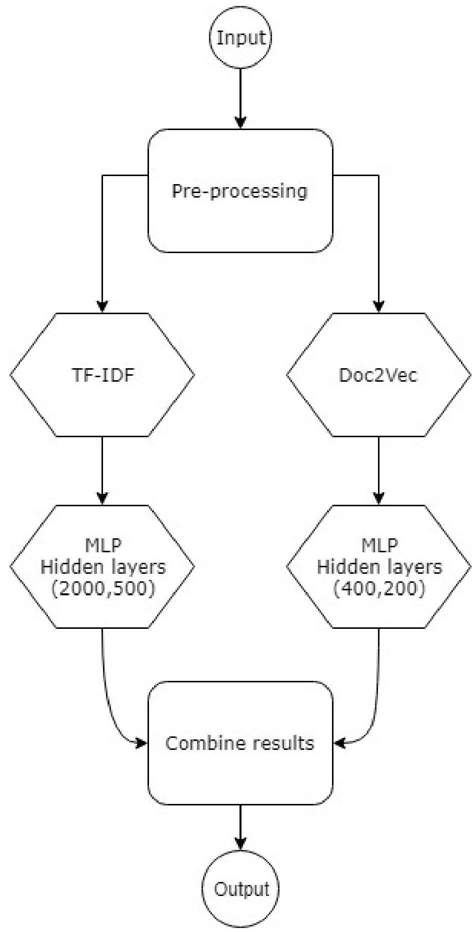

Figure 1. A picture of the proposed system.

4.3 Prediction Based on Doc2Vec Vector Representations

The second model of the stack is using Doc2Vec [3] representations for each author tweets. A Doc2Vec model is built using [9] and trained on the training corpus. It infers a 300-dimensional vector. Similarly to the preceding model, a feed-forward neural network with 2 hidden layers is trained for prediction. The hyper-parameters of the network are shown in Table 2.

5 Experiments and Results

In order to tune the hyper-parameters of the models, several experiments were conducted. Both models of the final stacked system were tuned individually. During the

Table 1. TF-IDF model hyper-parameters

Parameter	Best fit
Batch size	32
Epoch	7
Dropout regularization	0.2
Activation function	relu
Optimizer	Adadelta
Hidden layers	(2000, 500)

Table 2. Doc2Vec model hyper-parameters

Parameter	Best fit
Batch size	32
Epoch	7
Dropout regularization	0.2
Activation function	relu
Optimizer	Adadelta
Hidden layers	(400, 200)

Table 3. Results for both models separately and combination of them

model	accuracy
TF–IDF	83.62
Doc2Vec	83.22
Stacked	85.96

experiments the hyper-parameters were obtained as shown in Tables 1 and 2. The unofficial final results of the two models, tuned separately are shown in Table 3, where both models are trained on the training part of the dataset, and evaluated on the development part of the dataset. The accuracy is calculated for the gender classification sub-task (bot, male or female).

We can see that both models achieved similar results, which led to another experiment with both models combined. The combination of TF–IDF and Doc2Vec outperformed both models separately, so that it was the obvious choice for the final system at PAN 2019 [2], Bots and Gender Profiling task [8].

Finally, both models of the submitted stacked system were trained on the whole part of the training dataset and the stack was tested on the official test set on the TIRA platform [6].

6 Conclusion and Future Work

This paper has presented a system for predicting bots and gender from twitter feeds for the Bots and Gender Profiling task [8] at PAN 2019 [2]. The submitted system is a stack of two models, where each model uses different vectorial representations for the Twitter feeds; however, both models uses multilayer perceptron neural network. For vector representations TF–IDF and Doc2Vec were used.

The submitted system only works for English language. Therefore, for future research the system should be improved so that it can work with another languages as well.

References

1. Ciccone, G., Sultan, A., Laporte, L., Egyed-Zsigmond, E., Alhamzeh, A., Granitzer, M.: Stacked Gender Prediction from Tweet Texts and Images—Notebook for PAN at CLEF 2018. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) CLEF 2018 Evaluation Labs and Workshop – Working Notes Papers, 10-14 September, Avignon, France. CEUR-WS.org (Sep 2018), http://ceur-ws.org/Vol-2125/
2. 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)
3. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32. pp. II–1188–II–1196. ICML'14, JMLR.org (2014), http://dl.acm.org/citation.cfm?id=3044805.3045025
4. Loper, E., Bird, S.: Nltk: The natural language toolkit. In: In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics (2002)
5. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
6. 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)
7. Poulston, A., Waseem, Z., Stevenson, M.: Using TF-IDF n-gram and Word Embedding Cluster Ensembles for Author Profiling—Notebook for PAN at CLEF 2017. In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers, 11-14 September, Dublin, Ireland. CEUR-WS.org (Sep 2017), http://ceur-ws.org/Vol-1866/
8. 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)
9. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45–50. ELRA, Valletta, Malta (May 2010), http://is.muni.cz/publication/884893/en