Bot and Gender Detection using Textual and Stylistic Information

Notebook for PAN at CLEF 2019

Anastasia Giachanou and Bilal Ghanem

PRHLT Research Center, Universitat Politècnica de València, Spain
angia9@upv.es, bigha@doctor.upv.es

Abstract In this paper, we report our participation in the Bots and Gender Profiling task at PAN 2019. Our methodology on bots detection takes advantage of the tweets' textual and stylistic information. For the bots detection task, we propose a linear Support Vector Machine (SVM) classifier trained on words and character grams and on additional features that represent the variation in the use of the different stylistic features. For gender identification, we propose a Stochastic Gradient Descent (SGD) classifier trained on words and character grams, sentiment of tweets and the Pointwise Mutual Information (PMI) of terms. The PMI features represent the importance of the terms per gender. We have managed to achieve average accuracy scores of 0.881 and 0.7105 for the bot and gender identification tasks respectively.

Keywords: bot detection, gender detection, pointwise mutual information, social media

1 Introduction

The rise of social media has changed the way that people communicate and interact. Social media have given the opportunity to people to post and share their thoughts and opinions on any topic. However, they also allow malicious accounts to post and propagate fake news that have negative effects on the society. One common way to create fake news are the social bots. Social bots are computer algorithms that exhibit human-like behavior and can generate content and interact with users. Although there are some bots that perform useful tasks, there is also a growing number of bots that perform malicious functions and which aim to emulate and alter users' behavior [6]. For example, some bots can negatively affect democratic political discussions and influence public opinion during a presidential campaign or can promote terrorist propaganda and recruitment [2].

Detecting the bot accounts in an automatic way is very important for preventing the dissemination of fake news. Additionally, knowing the personality traits of social media users can potentially help to understand the personality traits that make users more vulnerable to propagate fake news or even to generate fake news or rumours [18]. Author profiling focuses on understanding different characteristics of the users depending on the linguistic patterns they use.

In this paper we present our system for bot and gender detection shared task at PAN 2019 [16,5]. The task focuses on investigating whether the author of a Twitter feed is a bot or a human, and in case of human, profiling the gender of the author. The task includes tweets in English and Spanish. For the bot detection task, we hypothesise that bots tend to post tweets with the same number of stylistic features (e.g., links, mentions). Therefore, we propose features that represent the standard deviation in the use of stylistic features. For the gender detection task, we assume that there are words that are more common in the tweets posted by male and words that are more common in the tweets posted by female users. Therefore, we propose using the Pointwise Mutual Information (PMI) of terms to assign a weight score to each term by gender class. Additionally, we use the emotion and sentiment scores expressed in the tweets assuming that different emotions are expressed in the tweets posted by male and female users. Both the bot and gender detection tasks are defined as binary classification tasks.

2 Related Work

Bot and gender detection have attracted a lot of research attention in the last years. Varol et al. [19] proposed a machine learning system that extracts features from six different categories: users and friends meta-data, tweet content and sentiment, network patterns, and activity time series to differentiate between bot and human accounts. Cai et al. [3] proposed a behavior enhanced deep model for bot detection. They proposed to extract latent temporal patterns based on user content. Additionally, they proposed to combine content and behavior information using deep learning method. A great number of researchers have focused on credibility and fake news detection. For example, Giachanou et al. [7] proposed EmoCred that incorporates emotions that are expressed in the claims into an LSTM network to differentiate between fake and real claims.

Gender detection has also attracted a lot of research attention. Schler et al. [17] performed gender classification on a corpus of 71,000 blogs using unigrams with the highest information gain and stylistic features. The results showed that male bloggers write more about politics and technology, while female bloggers write more about their personal lives. Rangel and Rosso [15] proposed the EmoGraph approach to capture how users convey verbal emotions in the morphosyntactic structure of the discourse.

3 Bots and Gender Detection Systems

First, we perform some preprocessing that is the same for both the bot and gender detection. The preprocessing includes the following steps:

- Concatenate all 100 tweets of each author into one document
- Replace URLs with the tag <URL>
- Replace the mentions with the tag <MENTION>
- Replace the symbol of # with the tag <HASHTAG>
- Lowercase all the characters

Here, we should mention that we did not remove the stopwords for any of the tasks.

3.1 Bots Detection System

Intuitively, content is the most important for the bots prediction task. To this end, our systems start with word grams and chargrams. Although, word grams and chargrams are simple features, they have been shown to be important for various tasks such as emotional reactions prediction [8,9] as well as information retrieval tasks [1,11]. In our bot detection system we use word grams that range from 1 to 3 and chargrams that range from 2 to 6.

In addition, we define some new features that capture how much the stylistic features of tweets posted by the same user vary. The intuition is that the bots will post tweets that are similar in terms of the number of each stylistic feature they contain (e.g., hashtags, mentions, exclamation marks), whereas the humans' tweets will be more diverse. More formally, let $T = \{t_1, ..., t_i, ..., t_{|T|}\}$ be the list of the tweets posted by a user. Also, let $V = \{v_1, ..., v_i, ..., v_{|V|}\}$ be a list of different stylistic variations and $c(t, v)$ the number of occurrences of stylistic variation v in a tweet t. Then we can calculate the deviations in the occurrences as:

$$SD_v = \sqrt{\frac{|c(t, v) - \mu_v|^2}{T}}$$

where μ_v is the mean of the occurrences of the variation v. We calculate all the deviations in a similar way. We use the deviations of the following stylistic variations: *exclamation marks, question marks, negative emoticons, positive emoticons, terms in capital, terms with repeated characters, mentions, links, hashtags*. In addition, we count the number of duplicate tweets.

We train a SVM classifier on the proposed features for the prediction in the bots detection system.

3.2 Gender Detection System

For the gender detection task, we learn the weights of the words for each gender class. To learn those weights, we use the Pointwise Mutual Information (PMI) method originally proposed by Church and Hanks [4]. According to this approach, every term w is assigned a PMI score for each of the two gender classes: male and female. The PMI score for a term w regarding the *male* class is calculated as follows:

$$PMI(w, male) = \log_2 \frac{c(w, male) * N}{c(w) * c(male)}$$

where $c(w, male)$ is the frequency of the term w in the tweets posted by a male user, N is the total number of words in the corpus, $c(w)$ is the frequency of the term in the corpus and $c(male)$ is the number of terms in the tweets posted by male users. The PMI of the terms for the female class is calculated in a similar way. Then the total PMI score for a document d regarding the male class can be calculated as:

$$PMI(d, male) = \sum_{w \in d} PMI(w, male)$$

Then these scores are used as additional to the word grams and chargrams features during the training phase. For English gender detection we use word grams that range from 1 to 3 and chargrams that range from 3 to 4, whereas for Spanish we use word grams from 1 to 2 and chargrams from 2 to 6. In addition, for every document d we calculate the emotion and sentiment scores. We follow a lexicon based approach to calculate these scores. More specifically, we simply count the number of occurrences of the emotional and sentimental words that occur in a document d. For the tweets that are written in English, we focus on *positive, negative, anger, anticipation, disgust, fear, joy, surprise, trust* and *sadness* whereas for the Spanish we focus only on *positive* and *negative*. Finally, we count the number of emoticons that appear in the text. We train a SGD classifier on the proposed features for the gender prediction task.

4 Experimental Setup

In this section we present the dataset and the experimental settings of our methodology.

4.1 Dataset

The dataset consists of tweets in English and Spanish. Table 1 shows the statistics on the dataset.

Table 1. Statistics of the dataset.

	Training		Development	
	Bot	Gender	Bot	Gender
English	2880	1440	1240	620
Spanish	2080	1040	920	460

We observe that the English collection is bigger than the Spanish. Here we should mention that the dataset is balanced over the classes for both bot and gender detection. For each user, a total of 100 tweets are provided.

4.2 Experimental Settings

The submission of our system was made from the TIRA platform [13]. We examined different classifiers including Logistic Regression, Random Forest, Support Vector Machine and Stochastic Gradient Descent. Regarding the bot detection task we obtained the best results with a linear SVM and regarding gender detection with SGD. Regarding SVM, we set the penalty parameter C to 10. For both SVM and SGD we use the square hinge loss. For the implementation of our system we use scikit-learn library[1]. To create our word and char grams we use the hashing vectorizer provided by scikit-learn library.

[1] https://scikit-learn.org

As already mentioned, on the gender detection task we use the emotions expressed in the text as additional features. For the English tweets we use the NRC-Emotion-Lexicon [10] whereas for the Spanish tweets we use the Spanish lexicon presented by Perez-Rosas et al. [12]. The submitted systems are compared with several deadlines including majority, random, char grams, word grams, word embeddings and the Low Dimensionality Statistical Embedding approach [14].

5 Results

Table 2 presents the results of our system on the bot and gender detection tasks on English and Spanish in terms of accuracy on the development set. We use the same word and char gram ranges that were used for our final system. Also, we use SVM and SGD for the bot and gender detection respectively. From the results we observe that for the bot detection task the best results are achieved with the combination of word grams and chargrams. For the gender detection the best results are achieved with the char-grams.

Table 2. Accuracy scores for word and char grams on the development set.

	Bot-English	Gender-English	Bot-Spanish	Gender-Spanish
char grams	0.935	0.790	0.911	0.687
word grams	0.927	0.793	0.908	0.674
word & char grams	0.933	0.759	0.912	0.667

Table 3 presents the results of different combinations of features on the development set regarding the bot detection task. We observe that the best performance is achieved using the word and char grams. Despite this result, we decided to use all the combination of all the features for our final system to examine their usefulness on the task.

Table 3. Accuracy scores for different features combinations on the development set on the bot detection task.

	Bot-English	Bot-Spanish
words & char grams	0.933	0.912
words & char grams & deviations	0.933	0.883
words & char grams & duplicates	0.931	0.876
words & char grams & deviations & duplicates	0.925	0.877

Table 4 presents the results of different combinations of features on the development set regarding the gender detection task. We observe that for the English gender detection the best result is achieved when all the features are combined. Regarding the Spanish gender detection, the best result is achieved with the word and char grams. For our final system, we decided to use the combination of all the features.

Table 4. Accuracy scores for different features combinations on the development set on the gender detection task.

	Gender-English	Gender-Spanish
words & char grams	0.759	0.667
words & char grams & PMI	0.760	0.584
words & char grams & PMI & sentiments	0.785	0.642

Table 5 presents the results of our system on the bot and gender detection tasks on English and Spanish in terms of accuracy on the official test set. We observe that our system performs better for the English tweets compared to Spanish. Also, we observe that the performance is higher for the bot detection task compared to the gender detection task.

Table 5. Accuracy scores of our system on the official test set.

	Bot	Gender
English	0.906	0.773
Spanish	0.856	0.648

6 Conclusions

In this paper we described our system for bot and gender detection task at PAN 2019. Regarding the bot detection we proposed a system trained on textual and stylistic features, whereas for the gender detection we proposed a system based on word and char grams, PMI weights of terms and sentiment features.

Our results showed that words and char grams are very important features for the bot and gender detection. Also, we showed that our system that was trained on the combination of all the proposed features managed to achieve 0.906 and 0.856 accuracy scores for the English and Spanish bot detection respectively. In addition, from our results we observe that our system on the bot detection performs better compared to our system on the gender detection.

Acknowledgments.

The first author is supported by the SNSF Early Postdoc Mobility grant P2TIP2_181441 under the project *Early Fake News Detection on Social Media*, Switzerland.

References

1. Aliannejadi, M., Crestani, F.: Venue suggestion using social-centric scores. CoRR abs/1803.08354 (2018)

2. Berger, J.M., Morgan, J.: The isis twitter census: Defining and describing the population of isis supporters on twitter. The Brookings Project on US Relations with the Islamic World 3(20), 4–1 (2015)
3. Cai, C., Li, L., Zengi, D.: Behavior enhanced deep bot detection in social media. In: Proceedings of the 2017 IEEE International Conference on Intelligence and Security Informatics. pp. 128–130. ISI '17 (2017)
4. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. Computational Linguistics 16(1), 22–29 (1990)
5. 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: Proceedings of the Tenth International Conference of the CLEF Association (CLEF 2019). Springer (2019)
6. Ferrara, E., Varol, O., Davis, C., Menczer, F., Flammini, A.: The rise of social bots. Communications of the ACM 59(7), 96–104 (2016)
7. Giachanou, A., Rosso, P., Crestani, F.: Leveraging emotional signals for credibility detection. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '19 (2019)
8. Giachanou, A., Rosso, P., Mele, I., Crestani, F.: Early commenting features for emotional reactions prediction. In: Proceedings of the International Symposium on String Processing and Information Retrieval. pp. 168–182. SPIRE '18 (2018)
9. Giachanou, A., Rosso, P., Mele, I., Crestani, F.: Emotional influence prediction of news posts. In: Proceedings of the 12th International AAAI Conference on Web and Social Media. pp. 592–595. ICWSM '18 (2018)
10. Mohammad, S.M., Turney, P.D.: Crowdsourcing a word-emotion association lexicon. Computational Intelligence 29(3), 436–465 (2013)
11. Paltoglou, G., Giachanou, A.: Opinion Retrieval: Searching for Opinions in Social Media, pp. 193–214 (2014)
12. Perez-Rosas, V., Banea, C., Mihalcea, R.: Learning sentiment lexicons in spanish. In: Proceedings of the 8th International Conference on Language Resources and Evaluation. pp. 3077–3081. LREC '12 (2012)
13. 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)
14. Rangel, F., Franco-Salvador, M., Rosso, P.: A low dimensionality representation forÂălanguage variety identification. In: Computational Linguistics and Intelligent Text Processing. pp. 156–169. CICLing 2016 (2018)
15. Rangel, F., Rosso, P.: On the impact of emotions on author profiling. Information processing & management 52(1), 73–92 (2016)
16. Rangel, F., Rosso, P.: Overview of the 7th Author Profiling Task at PAN 2019: Bots and Gender Profiling. In: CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2019)
17. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.W.: Effects of age and gender on blogging. In: AAAI spring symposium: Computational approaches to analyzing weblogs. vol. 6, pp. 199–205 (2006)
18. Shu, K., Wang, S., Liu, H.: Understanding user profiles on social media for fake news detection. In: Proceedings of the IEEE 1st Conference on Multimedia Information Processing and Retrieval. pp. 430–435 (2018)
19. Varol, O., Ferrara, E., Davis, C.A., Menczer, F., Flammini, A.: Online human-bot interactions: Detection, estimation, and characterization. In: Proceedings of the 11th International AAAI Conference on Web and Social Media. ICWSM (2017)