Bot and Gender Identification in Twitter using Word and Character N-Grams

Notebook for PAN at CLEF 2019

Inna Vogel and Peter Jiang

Fraunhofer Institute for Secure Information Technology SIT
Rheinstrasse 75, 64295 Darmstadt, Germany
Inna.Vogel@SIT.Fraunhofer.de

Abstract Automated social media accounts, called bots, gained worldwide considerable importance over the course of the last years. Social bots can have serious implications on our society by swaying political elections or spreading disinformation - giving rationale to social bot detection as an emerging research area. Hence, tools and techniques to automatically detect and classify manipulative bots are needed.

In this notebook, we describe our system for the author profiling task at PAN 2019 on bot and gender identification on Twitter. The submitted system uses word unigrams and bigrams as well as character n-grams as features. Tweet preprocessing and feature construction were conducted to train a linear Support Vector Machine (SVM) classifier. Our model shows that it is possible to differentiate bots from humans with a (fairly) high accuracy. Additionally, the accuracy shows that our SVM architecture can solidly determine the gender of the author (male or female). Our submitted model achieved an overall accuracy of 0.92 for bot detection on the English dataset and an accuracy of 0.91 for Spanish tweets. Gender can be determined by the accuracy of 0.82 and 0.78 on the English and Spanish corpus, respectively. Our simple model ranked 8th out of 55 competitors.

Keywords: Author Profiling, Bot Detection, Gender Detection, Natural Language Processing, Twitter

1 Introduction

A social bot, short for software robot, is a computer algorithm that is able to automatically produce content and interact with humans on social media. Useful bots, for instance, can automatically aggregate content from various sources to generate simple news feeds. However, due to recent reports of social media manipulation, including disinformation and extremism, concerns about the abuse of social bots are increasing

[7]. Some bots are very simple and solely retweet posts, whereas others are sophisticated and have the capability to interact with humans [3]. By emulating the behavior of humans, social bots have been used to infiltrate political discourse, manipulate the stock market and spread misinformation, rumors, spam, slander, or even just noise [6]. Therefore, the detection of social bots is an important research area.

Author profiling uses information that is for example available on social media platforms to determine various characteristics of an author, for instance, through the analysis of his or her documents. These traits can be the author's gender, age, personality, or the cultural and social context, such as native language and dialects [19]. Author profiling is not only used in areas like marketing but also as a valuable additional tool in criminal investigations and in security [16].

This year, the author profiling task [14] of PAN 2019 [1] [4] was conceived to investigate whether the author of a Twitter feed is a bot or a human. If the author was decided to be human, an additional task was to infer its gender - either male or female. This year's dataset, which was provided by the PAN organizers, includes two languages: English and Spanish.

In the following chapters, we describe our approach in bot and gender identification in the author profiling task at PAN 2019. First, we summarize the related work (section 2). In section 3, we describe the Twitter data that was provided by the PAN organizers for training purpose. The preprocessing steps and features used to train a linear Support Vector Machine (SVM) are detailed in section 4. We consider the bot and gender profiling challenge as a multi-class, rather than a binary classification problem. Section 5 reports the accuracy scores of our model on the official PAN test dataset, whereas in section 6, we give a review of alternatively tested approaches. Section 7 concludes our work with a brief discussion.

2 Related Work

Bot Classification. Gilani at al. [8] analyzed behavioral characteristics of bots and humans in Twitter data, such as likes, retweets, user replies and mentions, activity or the size of the uploaded content. They also counted the shared URLs and the follower-friend ratio of each Twitter user account. For data collection, preprocessing, annotation and analysis the framework known as *Stweeler* [9] was used. The authors [8] observed that humans generate more novel content, whereas bots rely on retweeting. Also, bots have a higher propensity to share URLs and upload media (such as images and videos) more frequently than humans.

Varol et al. [20] introduced a Twitter bot detection framework that extracts more than a thousand features from six different classes of Twitter users data and metadata. Extracted features are, for example, the number of friends of the Twitter user, the tweeted content, sentiment found in the text or his or her activity time series. The extracted features were then used to train different models in order to detect social bots. By using a 5-fold cross-validation procedure, the trained Random Forest classifier achieved an accuracy of 0.95 AUC.

[1] PAN at CLEF 2019 Bots and Gender Profiling: https://pan.webis.de/clef19/pan19-web/author-profiling.html

Gender Detection. While in 2017 the objective was to identify gender and language variety in twitter data in four languages (English, Spanish, Arabic, and Portuguese) [17], the focus of the 2018 task was to address gender identification from a multimodal perspective. For this, not only texts, but also images were given for the classification task [15].

Rao at al. [18] performed gender detection on tweets written in English using a Support Vector Machine (SVM). By combining n-grams and sociolinguistic features like emoticons or alphabetic character repetitions, they were able to achieve an accuracy of 0.72. Basile et al. [2] also trained a SVM with the combination of character and tf-idf n-grams and report an overall accuracy of 0.83. Following another approach, Martinc et al. [12] achieved an accuracy of 0.82 by training a Logistic Regression classifier in combination with a variety of features like character, word, and POS n-grams, emoticons, sentiments, character flooding and list of words.

Besides linguistic features, researchers have also analyzed language independent features to predict the gender of users. Alowibdi et al. [1] presented a novel approach by utilizing color-based features extracted from the profile layout colors (i.e. background) on Twitter. They claim their approach to be scalable, efficient and to have low computational complexity. [1] applied four different classifiers, namely Probabilistic Neural Network (PNN), Decision Tree (DT), Naïve Bayes (NB) and Naïve Bayes / Decision-Tree Hybrid (NB-Tree), each with 10-fold cross-validation. They stated that NB-Tree performed with an accuracy of around 0.70 consistently in all their experiments.

3 Dataset Description

The provided training dataset of the bots and gender profiling task [14] at PAN 2019 consists of 4,120 English and 3,000 Spanish Twitter user accounts. Each of these XML files contains 100 tweets per author. Every tweet is stored in a <*document*> XML tag. Figure 1 shows the dataset structure for bot and gender detection.

Every author was coded with an alphanumeric author-ID. The English language folder contains 2,060 bot texts, 1,030 female and the same amount of male texts. The Spanish folder is smaller than the English one and includes 1,500 bot texts and 750 texts

Figure 1. Dataset Structure of the PANs 2019 Bots and Gender Profiling Task

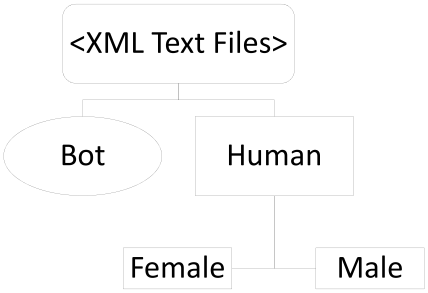

Table 1. Train and Test Split of the Bots and Gender Profiling Dataset as recommended

	English (EN)		**Spanish (ES)**	
	Bot	Gender (Female & Male)	Bot	Gender (Female & Male)
Training	1,440	720 x 2	1,040	520 x 2
Test	620	310 x 2	460	230 x 2
Total	2,060	2,060	1,500	1,500

per gender. To avoid overfitting while training a classifier, the data is split into a training (70%) and testing (30%) set - as recommended by the PAN organizers (see Table 1).

When considering the binary classification problem ”Bot vs Human“, the dataset is balanced. But if it is reformulated to a three-class problem, the ”Bot“ class predominates the two gender classes- ”Male“ and ”Female“. Unbalanced data refers to an unequal distribution of class instances. This imbalance can be reduced to a great extent by employing the ”Undersampling“ technique. By randomly removing samples from the majority class, this simple method allows creating balanced datasets that, in theory, result in classifiers that are not biased towards a certain class. By undersampling the bot class we took the risk leaving out important instances that may provide important differences between the three classes.

The number of English bot texts was reduced from 2,060 to 1,030, according to the size of male and female authors per class. The Spanish bot texts were reduced from 1,500 to 750 instances. Additionally, we partitioned the training dataset into three smaller sets. 50% of the data was used for training, 25% for tuning and 25% for testing the performance of the SVM. Following common practice, the validation split was used to optimize the parameters of the model.

The gold-standard was only accessed through the TIRA [13] evaluation platform provided by the PAN organizers.

4 Methodology

In the following, the same approach is applied for each language. First, we preprocess the Twitter data to handle idiosyncrasies such as hashtags, URLs and user mentions. Afterwards, word unigrams, and bigrams, as well as character n-grams in the range from 3 to 5 are extracted as features which serve as input to train a Support Vector Machine (SVM).

4.1 Preprocessing

The preprocessing pipeline is for both languages (English and Spanish) nearly the same. The following steps are performed to clean and structure the Tweets:

1. Concatenation of all 100 tweets per author into one string.
2. Lowercasing all characters.
3. Removing white space.
4. Replacing URLs with the placeholder <*URL*>.

5. Deleting irrelevant signs, e.g. *”+,*,/,“*.
6. Replacing all hashtags and attached tokens with the placeholder <*HashTag*>.
7. Replacing user mentions (e.g. @username) with the placeholder <*UsernameMention*>.
8. Sequences of the same characters with a length greater than three are removed.
9. Removing words with less than three characters.
10. Removing stopwords by using the NLTK (Natural Language Toolkit) list of stopwords for the English and Spanish language.
11. To tokenize the words we used the *TwitterTokenizer* from the NLTK library. The TwitterTokenizer is adapted for Twitter and other forms of casual speech used in social networks. It contains some regularization and normalization features (e.g. converting tweets to lowercase and vice-versa, removing username mentions and reducing the length of words in the tweet with repeated characters).

4.2 Features

Since the two languages process different datasets, two separate classification models for each language were trained. We tested different feature sets and experimented with hyper-parameter tuning, manually and by employing scikit-learn’s grid search function. The hyper-parameters were tuned for each language model separately. Different experiments are discussed in Section 6.

After preprocessing, a term frequency distribution (tf) on the two datasets was performed. We concatenated the training, validation and test sets. The three most frequently used tokens by bots are:

a) URLs (token <*URL*>)
b) Hash Tags (token <*HashTag*>)
c) and @-mentions (token <*UsernameMention*>)

While bots have a higher propensity to share URLs, humans tend to primarily refer to other users (or accounts) by using @-mentions (marked as <*UsernameMention*> token during preprocessing). Besides referring to other users, humans share URLs and use hashtags (identified as second and third most used token). This analysis shows that special attention should be given to these tokens when Twitter texts are preprocessed.

According to the frequency distribution, the 10,000 most frequently used tokens in the training set were stored in a dictionary. When building the vocabulary, terms that have a document frequency lower than 2 were ignored.

We used scikit-learn’s term frequency-inverse document frequency (tf-idf) weighting function (*tfidfVectorizer*) to convert the tokens to a matrix of tf-idf features in order to build a vector pipeline for each language. The following n-gram features for both languages were used:

a) word unigrams and bigrams
b) character n-grams in the range 3 to 5

The way the word and character selection was conducted is inspired by Daneshvar and Inkpen [5]. The authors presented their gender identification approach for Twitter texts at the author profiling shared task at PAN in 2018 in which their model was ranked second.

4.3 Machine Learning Algorithm

To train a classifier we used a linear SVM with different types of word and character n-grams as features. As we consider the task a multi-class classification problem, the OVR (”One-vs.-Rest“) decision function was used. OVR combines multiple binary SVMs to solve the multi-class classification task with the training of multiple number of classes [11]. Our three classes to train are: ”Bot“, ”Male“ and ”Female“. Using OVR each SVM classifies samples into the corresponding class against all the others [10].

In order to avoid overfitting when experimenting with the training set, we split the data provided by the organizers in three parts. For training, we used 50% of the data. The other half of the dataset was divided equally as validation and test set (each 25% of the text data). During the experiments, the model did not see the test set. Parameter tuning was performed on the validation data set. Finally, each model was tested on the official PAN 2019 test set for the Author Profiling task on the TIRA platform [13]. The classification results are provided in the following section.

5 Results

The performance of the author profiling task was ranked by accuracy. Table 2 shows the scores for our final system performed on the ”early bird“ training set, as well as the accuracy scores on the official test set. Accuracy scores were calculated individually for each language. First, the accuracy for identifying bot vs. human was calculated. Then, in the case of a human, the accuracy of predicting the human as male or female was calculated. Each model was trained on 50% of the test data. Hyper parameters were tuned on the 25% validation split. Finally, the submitted model was tested on the official PAN 2019 test set on the TIRA platform. The results are displayed in Table 2.

Table 2. Accuracy Scores for Bot and Gender Detection on the ”Early Bird“ and Official PAN 2019 Test Dataset

	”Early Bird“ Dataset		**Test Data Set**	
Language	EN	ES	EN	ES
Bot vs. Human	0.97	0.97	**0.92**	**0.91**
Male vs. Female	0.94	0.93	**0.82**	**0.78**

6 Other Tested Methods and Features

Next to the preprocessing and feature selection steps already described, we also examined other features and data structure techniques. Apart from the presented SVM with a linear kernel, we also tested other classifiers, namely CNN and the Random Forest Classifier. In our experiments these two classifiers were not able to keep up with the linear SVM classifier in terms of performance.

In our first experiments, Twitter data was cleaned by removing URLs, hashtags, retweets (RT) and user mentions. Experiments have shown that these features are essential for the bot detection task on Twitter data. To vectorize our tokens, we first worked with total and relative word frequencies as well as with converting the tokens to tf-idf features. The vector length ranged between 1,000 and 10,000 most frequent tokens. Our experiments showed that the accuracy dropped when the hyper-parameters were tuned using scikit-learn's grid search function. Table 3 shows the results of our experiments tested on the ”early bird“ test dataset.

Table 3. Accuracy Scores for Bot and Gender Detection Experiments on the PAN 2019 ”Early Bird“ Test Dataset

	Bot vs. Human		Male vs. Female	
	EN	ES	EN	ES
Token Total Frequency	0.91	0.83	0.65	0.64
Token Relative Frequency (Grid Search Tuning)	0.73	0.83	0.53	0.64
Token TF-IDF Vectorization	0.92	0.78	0.81	0.61

7 Discussion and Conclusion

In this notebook, we described our system for the author profiling task at PAN 2019 on bot and gender detection. A SVM was trained to classify a text produced either by a bot or a human. And in case of a human, whether the text was written by a male or female author. By using word unigrams and bigrams as well as character n-grams as features, our model demonstrated that computer algorithms, which are able to automatically produce content, are detectable with high accuracy.

Since we considered the author profiling challenge as a multi-class, rather than a binary classification problem, we restructured the provided Twitter data by undersampling the bot data. In doing so, we created three balanced datasets (bot, male and female), avoiding the classifier being biased towards the ”Bot“ class.

As far as the preprocessing of raw data is concerned, it is worth to mention that special attention should be paid to URLs, hashtags and user mentions, as these features can help to improve the performance of a classifier. After performing a term frequency distribution on the English and Spanish dataset each, the assumption that bots have a higher propensity to share URLs, while humans tend to primarily refer to other users by using @-mentions, was confirmed.

We experimented with different features, vectorization techniques and dimensionality sizes. Our model performed best by using 10,000 most common word unigrams and bigrams as well as character n-grams in the range of 3 to 5 with official final scores of 0.92 for English and 0.91 for Spanish bot detection. With regard to gender recognition, an accuracy of 0.82 and 0.78, respectively, could be achieved for the English and Spanish data sets. Our simple approach ranked 8th out of 55 competitors.

Acknowledgements

This work was supported by the German Federal Ministry of Education and Research (BMBF) under the project “DORIAN” (Scrutinise and thwart disinformation).

References

1. Alowibdi, J.S., Buy, U.A., Yu, P.: Language independent gender classification on twitter. In: Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining. pp. 739–743. ACM (2013)
2. Basile, A., Dwyer, G., Medvedeva, M., Rawee, J., Haagsma, H., Nissim, M.: Is there life beyond n-grams? a simple svm-based author profiling system. In: Working Notes of CLEF 2017-Conference and Labs of the Evaluation Forum (2017)
3. Boshmaf, Y., Muslukhov, I., Beznosov, K., Ripeanu, M.: Design and analysis of a social botnet. Computer Networks 57(2), 556–578 (2013)
4. 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)
5. Daneshvar, S., Inkpen, D.: Gender identification in twitter using n-grams and lsa: Notebook for pan at clef 2018. In: CLEF (2018)
6. Ferrara, E., Varol, O., Davis, C., Menczer, F., Flammini, A.: The rise of social bots. Commun. ACM 59(7), 96–104 (Jun 2016), http://doi.acm.org/10.1145/2818717
7. Gayo-Avello, D.: Social media won’t free us. IEEE Internet Computing 21(4), 98–101 (2017)
8. Gilani, Z., Farahbakhsh, R., Tyson, G., Wang, L., Crowcroft, J.: Of bots and humans (on twitter). In: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017. pp. 349–354. ACM (2017)
9. Gilani, Z., Wang, L., Crowcroft, J., Almeida, M., Farahbakhsh, R.: Stweeler: A framework for twitter bot analysis. In: Proceedings of the 25th International Conference Companion on World Wide Web. pp. 37–38. International World Wide Web Conferences Steering Committee (2016)
10. Hong, J.H., Cho, S.B.: Multi-class cancer classification with ovr-support vector machines selected by naïve bayes classifier. In: King, I., Wang, J., Chan, L.W., Wang, D. (eds.) Neural Information Processing. pp. 155–164. Springer Berlin Heidelberg, Berlin, Heidelberg (2006)
11. Huang, D., Zhang, X., Huang, G.: Advances in Intelligent Computing: International Conference on Intelligent Computing, ICIC 2005, Hefei, China, August 23-26, 2005, Proceedings. No. Teil 1 in Lecture Notes in Computer Science, Springer Berlin Heidelberg (2005), https://books.google.de/books?id=ZiAHCAAAQBAJ
12. Martinc, M., Skrjanec, I., Zupan, K., Pollak, S.: Pan 2017: Author profiling-gender and language variety prediction. In: CLEF (Working Notes) (2017)
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., 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)

15. Rangel, F., Rosso, P., Montes-y Gómez, M., Potthast, M., Stein, B.: Overview of the 6th author profiling task at pan 2018: multimodal gender identification in twitter. Working Notes Papers of the CLEF (2018)
16. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the author profiling task at pan 2013. In: CLEF Conference on Multilingual and Multimodal Information Access Evaluation. pp. 352–365. CELCT (2013)
17. 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)
18. Rao, D., Yarowsky, D., Shreevats, A., Gupta, M.: Classifying latent user attributes in twitter. In: Proceedings of the 2nd international workshop on Search and mining user-generated contents. pp. 37–44. ACM (2010)
19. Russell, C.A., Miller, B.H.: Profile of a terrorist. Studies in conflict & terrorism 1(1), 17–34 (1977)
20. Varol, O., Ferrara, E., Davis, C.A., Menczer, F., Flammini, A.: Online human-bot interactions: Detection, estimation, and characterization. In: Eleventh international AAAI conference on web and social media (2017)