Bots and Gender Profiling on Twitter using Sociolinguistic Features

Notebook for PAN at CLEF 2019

Edwin Puertas[2,1,3], Luis Gabriel Moreno-Sandoval[1,3], Flor Miriam Plaza-del-Arco[4], Jorge Andres Alvarado-Valencia[1,3], Alexandra Pomares-Quimbaya[1,3], and L.Alfonso Ureña-López[4]

[1] Pontificia Universidad Javeriana, Bogotá, Colombia
{edwin.puertas,jorge.alavarado,morenoluis,pomares}@javeriana.edu.co
[2] Universidad Tecnológica de Bolívar, Cartagena, Colombia
epuerta@utb.edu.co
[3] Center of Excellence and Appropriation in Big Data and Data Analytics (CAOBA)
[4] Universidad de Jaén, Jaén, Andalucía, Spain.
{fmplaza, laurena}@ujaen.es

Abstract Unfortunately, in social networks, software bots or just bots are becoming more and more common because malicious people have seen their usefulness to spread false messages, spread rumors and even manipulate public opinion. Even though the text generated by users in social networks is a rich source of information that can be used to identify different aspects of its authors, not being able to recognize which users are truly humans and which are not, is a big drawback. In this work, we describe the properties of our multilingual classification model submitted for PAN2019 that is able to recognize bots from humans, and females from males. This solution extracted 18 features from the user's posts and applying a machine learning algorithm obtained good performance results.

Keywords: Bots profiling, gender profiling, author profiling, sociolinguistic, computational linguistic, user profiling

1 Introduction

Recent studies conducted by Yang [15] indicate that there is a steady growth of autonomous artificial entities known as social bots on digital platforms such as Twitter, which have allowed them to spread messages and influence large populations with ease. That study concludes in their research that between 9% and 15% of Twitter accounts show similar behaviors to bots [2,13,14].

Bots can be designed for doing malicious activities to manipulate opinions in a certain domain. These bots mislead, exploit, and manipulate social media discourse with rumors, malware, misinformation, spam, slander, among others [7]. Some emulate human behavior to enact fake political support or change the public perception of political entities [12]. For instance, social bots was distorted the 2016 U.S Presidential election online discussion, according to a report published by researchers at Oxford University[5]. Also, bots are used in the marketing area to manipulate the stock market [7] or terrorist purposes to promote terrorist propaganda and recruitment [1]. The detection of social bots is therefore an important research endeavor.

The automatic detection of bots in social media has attracted the attention of researchers in recent years. In fact, many techniques to analyze this problem are proposed in the literature. If we focus on systems based on feature-based machine learning methods, we found several works, such as the one proposed by David et al [5] that study the first social bot detection framework publicly available for Twitter. It analyzed more than 1000 features and grouped them into six classes: network, user, friends, temporal, content, and sentiment. On the other hand, Dickerson [6] proposed SentiBot, an architecture and associated set of algorithms that automatically identify bots on Twitter by using a combination of features including tweet sentiment and they conclude that a number of sentiment related factors are key to the identification of bots.

In this paper, the proposal is described as part of our participation in the Bots and Gender Profiling task of PAN 2019 [11,4] at CLEF. This task is focused on investigating whether the author of a Twitter account is a bot or a human. Furthermore, in case of human, to profile the gender of the author. For that purpose, we study the generations and analysis of different sociolinguistic features in order to identify how various linguistic characteristics differ between bots and humans and women and men.

The rest of the paper is structured as follows. In Section 2, we explain the data used in our methods. Section 3 presents the details of the proposed system. In Section 4, we discuss the analysis and evaluation results of our system. We conclude in Section 6 with remarks on future work.

2 Data Description

This year's task of author profiles of PAN 2019 is to predict if the author on Twitter is a bot or a human. The dataset contains tweets in English and Spanish as shown in Table 1. The data is split evenly between human and bot users. The tweets recovered for each user come from their timeline, which can vary between days and months depending on the frequency of use. Finally, the last 100 tweets from a timeline were recovered for each author.

3 Model Description

In this section, we explain the multilingual predictive model used in our submission. The model used for the task of Bots and Gender profiling in PAN 2019 [11,4], was designed to identify two types of classes: bot and gender. We proposed two hypotheses in accordance with the attributes of the dataset and the goals of the task, which are described in detail in Table 2.

[5] https://nyti.ms/2mNTwnk

According to the hypothesis presented in Table 2, we proposed two strategies. The first one generates features from the vocabulary terms used in the tweets. The second one, computes statistics for each profile to characterize the use of terms, hashtags, mentions, URLs, and emojis. On the basis of the proposed strategies, the "Training System" was designed. Figure 1 shows the proposed system to predict bot and gender, which consists of the following stages: preprocessing, standardization and transformation, extraction of features, configuration and classification, and testing.

3.1 Preprocessing

In the preprocessing stage, we use the concatenated vocabulary terms of each user's tweets, in order to have only one document per user profile. In addition, we applied the re-labeling of the hashtags using the word "label_hashtag", the mentions word with the word "label_mention", the URLs with the word "label_url", and the emojis by UTF-8 were replaced with the word "label_emoji". Finally, globally re-tagged words are searched and counted.

3.2 Normalization and Transformation

The next stage is associated with the normalization and transformation process. The normalization process generates random samples for the training and testing process.

During the transformation process, the vector representation of words is performed and the features for each user profile are calculated. This process can be configured in such a way that the vectorial representation of the words can be done with "N-gram" and the global features related to the tweets of the user profiles can also be parameterized.

It must be taken into account that the transformation process can be configured in such a way that the vectorial representation of the words can be done with "N-grams" and the global features related to the tweets of the user profiles can also be parameterized.

3.3 Feature Extraction

According to [8], human knowledge is distributed among a large number of information sources, with data volumes constantly growing. Social networks have become indispensable tools for the automatic understanding of language because they allow us to model the user's writing habits by extracting features from the texts published by them.

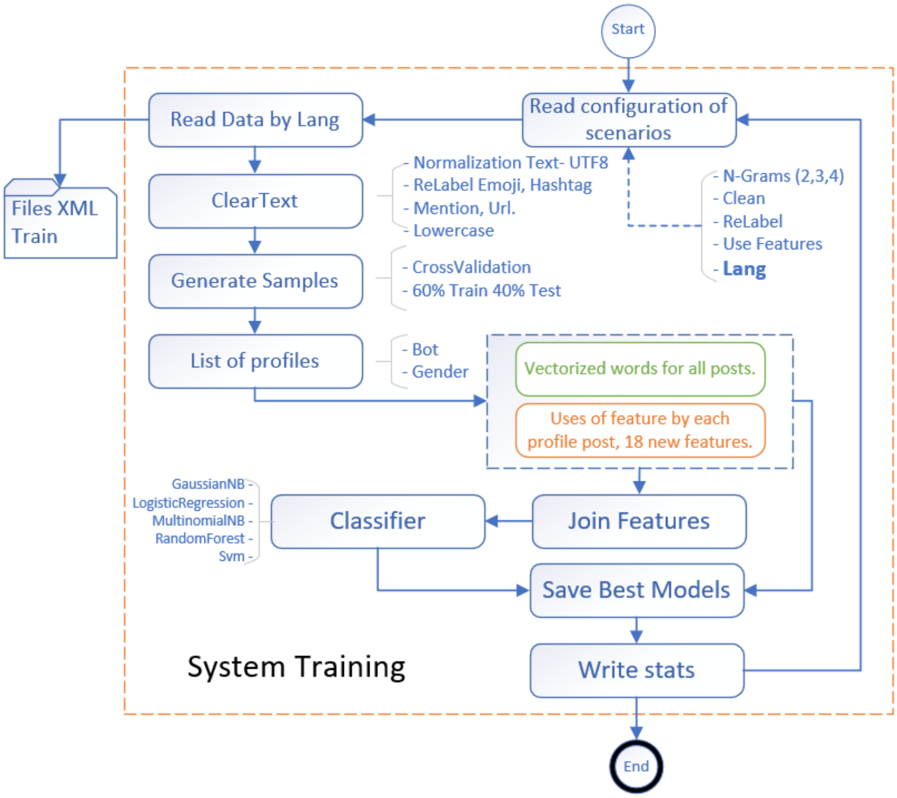
Table 1. Characteristics of training dataset.

Statistic	English	Spanish
# Bots User	4.120	1.500
# Human User	4.120	1.500
Avg. tweets per bot user	100	100
Avg. tweets per human user	100	100

Table 2. Description of hypothesis – H0

Class	Description - H0
Bots	For the hypothesis of bots classification, it is suggested that bots have less linguistic diversity than humans. For this reason, it was proposed to use classifiers that use vocabulary features and linguistic diversity.
Gender	For the hypothesis of gender classification, we believe that the vocabulary used by users can be associated with the use of linguistic features. For this reason, we analyze the way authors use emojis, hashtags, and mentions in addition to the vocabulary.

Figure 1. System Training.



In fact, the main challenge of the task for classifying Bots and Gender is associated with the detection of writing style on Twitter. According to [3], tweets produced by bots have a high amount of URLs compared to human tweets, thus, calculating the average of URLs per tweet is a valuable feature for classification algorithms. In addition,

Table 3. Features Description

#	Feature	Description
1	stats_avg_word	Average word size per tweet
2	stats_kur_word	Kurtosis of the variable stats_avg_word
3	stats_label_emoji	Amount of emojis per tweet for the profile
4	stats_label_hashtag	Number of hastags per tweet for the profile
5	stats_label_mention	Number of mentions per tweet for the profile
6	stats_label_url	Number of urls per tweet for the profile
7	stat_label_retweets	Number of retweets per tweet for the profile
8	stat_lexical_diversity	Lexicon diversity for all tweets by profile
9	stats_label_word	Number of words per tweet for the profile
10	kurtosis_avg_word	Kurtosis of the variable stats_kur_word
11	kurtosis_label_word	Kurtosis of the variable stats_label_word
12	skew_avg_word	Statistical asymmetry of the variable stats_avg_word
13	skew_label_word	Statistical asymmetry of the variable stats_avg_word
14	stats_person_1_sing	Number of tweets used by the first person of the singular
15	stats_person_2_sing	Number of tweets used by the second person singular
16	stats_person_3_sing	Number of tweets used by the third person singular
17	stats_person_1_plu	Number of tweets used by the first and second person of the plural
18	stats_person_3_plu	Number of tweets used by the third person plural

it is well known that people don't always spell words, hashtags, mentions, URLs and emojis correctly. For the aforementioned reasons, we extracted features at two levels: the tweet and the user profile level. At the tweet level we extracted the words, and the counts of hashtags, mentions, URLs, and emojis. At the user profile level we integrated the results obtained in the previous level calculating the average, kurtosis and asymmetry of the counts of hashtags, mentions, URLs, and emojis. Likewise, we analyze the lexical diversity comparing the words used in one tweet to the words used in the rest of the tweets.

3.4 Settings and classifiers

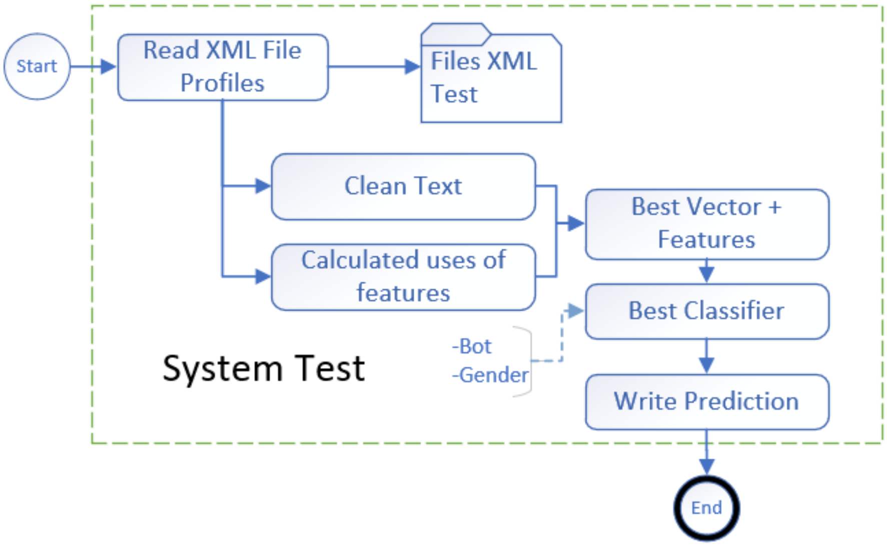
At the configuration stage, the system will adjust machine hardware parameters such as processors and threads. In addition, different scenarios can be configured for the use of the classifiers. Finally, the system may be adjusted to store the best performing vector words and qualifiers. It should be noted that during the execution of the system, the data set was divided into 60% for training and 40% for tests for all our experiments.

On the other hand, based on the goals of the task and on previous results of the author profiling tasks in the PAN, we analyzed different classifiers such as Naive Bayes (NB), Gaussian Naive Bayes (GNB), Complement Naive Bayes (CNB), Logistic Regression (LR), and Random Forests (RF).

3.5 Test

During the test stage, a software component was developed. It first reads the test data sets. Then the tweets are processed independently for each user profile. Afterwards,

Figure 2. System Training.



it calculates the features for each user. Subsequently, vector representation is made. The best classifiers for bots and gender classes are then calculated. Finally, the best predictors are exported. Figure 2 shows the "System Test" used by our models.

4 Experiments and Analysis of Results

During the pre-evaluation phase we carried out different experiments and the best ones were taken into account for the evaluation phase. The system was evaluated using the usual competition metrics, including Accuracy (Acc), Precision (P), Recall (R) and F1-score (F1). The best systems for bots and gender classification in the pre-evaluation phase will be explained in detail in the following sections.

It should be noted that the system presented was trained and tested with the dataset provided by the official site of PAN 2019 [4]. In addition, submissions were made on the TIRA platform [9] for the task of bots and gender profiles. The results obtained after evaluating our system with training dataset is shown in Table 4. The system uses various classification algorithms, such as Random Forest, GaussianNB, ComplementNB and Logistic Regression. But in the case of the English language Random Forest obtained better performance for bots and gender. And for the Spanish language Random Forest had better accuracy for Bots while Logistic Regression had better accuracy for genre.

Table 4. Summary of results in bots and gender classification per language

Type	Language	Acc	Best Model
BOT	en	0.91	RF
GENDER	en	0.81	RF
BOT	es	0.90	RF
GENDER	es	0.75	LR

Table 5. Bots classification in English and Spanish

	Precision		**Recall**		**F1-Score**		**Support**	
Class	en	es	en	es	en	es	en	es
0	0.97	0.96	0.85	0.82	0.91	0.88	620	460
1	0.86	0.84	0.98	0.97	0.92	0.90	620	460
Micro avg	0.91	0.89	0.91	0.89	0.91	0.89	1240	920
Macro avg	0.92	0.90	0.91	0.89	0.91	0.89	1240	920

Table 6. Gender classification in English and Spanish

	Precision		**Recall**		**F1-Score**		**Support**	
Class	en	es	en	es	en	es	en	es
0	0.79	0.76	0.84	0.72	0.81	0.74	310	540
1	0.83	0.73	0.77	0.78	0.80	0.76	310	540
Micro avg	0.81	0.75	0.81	0.75	0.81	0.75	620	1080
Macro avg	0.81	0.75	0.81	0.75	0.81	0.75	620	1080

4.1 Bots classification

Table 5 hows the results we have obtained for bots classification in English and Spanish languages after evaluating our system with training dataset. The best results were the Random Forest classifier for English language with 91% macro-F1 score and Spanish language with 89% macro-F1 score.

4.2 Gender classification

Table 6 hows the results we have obtained for gender classification in English and Spanish languages after evaluating our system with training dataset. The best results were the Random Forest classifier for English languages with 81% macro-F1 score and Logistic regression classifier for Spanish language with 75% macro-F1 score.

4.3 Submission Results

Table 7 shows the results for English and Spanish of the bots and the gender classification corresponding to the evaluation phase by means of a low dimensionality representation baseline for the identification of linguistic varieties (LDSE) [10]. For this, the

Table 7. Final classification

Dataset	training-dataset- 2019-02-18		test-dataset1- 2019-03-20		test-dataset2- 2019-04-29	
	es	en	es	en	es	en
Bot	0.84	0.91	0.70	0.90	0.81	0,88
Gender	0.80	0.84	0.61	0.78	0.69	0.76

different datasets provided by the task in that phase were applied. The measure used was the macro-F1 score, which was used to determine a weighted single value of the precision and integrity of the models used. It should be noted that the final results were obtained with the test2 dataset. In the general ranking of the task, we occupy the 33th position and we occupy the 9th position respected to baseline LDSE.

5 Discussion and Conclusion

The task of Bots and Gender profiling CLEF PAN 2019 [11,4] involved different tasks. The first one was the preprocessing of the corpus, which was composed of 100 posts for each user profile, for a total of 300.000 posts. Fortunately, the quality assurance during this preprocessing was not a challenge because the tweets were cleaned and the dataset balanced for each one of the target classes. On the contrary, feature extraction was one of the most significant challenges, because it was necessary to achieve a good performance with few samples of texts per user profile. To deal with this we decided to extract features at two levels: the tweet and the user profile. The first level aimed to obtain traditional counting values of words, hashtags, mentions, URLs, and emojis per tweet. The second one was intended to explore the author's diversity based on the analysis of the features extracted at the first level. The resulted features demonstrated to be very useful to discriminate bots from humans, and the different genders.

Regarding the second task, the classification itself, it was necessary to evaluate different techniques with different parametrization and different inputs. The final results demonstrated that Random Forest and Logistic Regression were the most relevant techniques for this problem.

In addition, during the final task, the evaluation of the model, we demonstrated our hypothesis, the lexical diversity, expressed using the 18 features, is a well discriminant for the target classes. It is important to highlight that for the classification of bots the best classifier using the n-grams and the proposed features obtained from the training dataset got an accuracy of 0.912, and using only the proposed features in the study it got 0.907 of accuracy. This demonstrates the predictive value of these features for the bots problem.

Finally, there are still issues to explore. One important aspect is to improve the profile analysis from the sociolinguistic point of view integrating features that describe the interaction dynamics of each user.

6 Acknowledgements

We thank the Center for Excellence and Appropriation in Big Data and Data Analytics (CAOBA), Pontificia Universidad Javeriana, and the Ministry of Information Technologies and Telecommunications of the Republic of Colombia (MinTIC). The models and results presented in this challenge contribute to the building of the research capabilities of CAOBA. Also, Fondo Europeo de Desarrollo Regional (FEDER) and REDES project (TIN2015-65136-C2-1-R) from the Spanish Government. Finally, the author Edwin Puertas gives thank to Universidad Tecnológica de Bolívar. Needless to say, we thank the organizing committee of PAN, especially Paolo Rosso, Francisco Rangel, Matti Wiegmann and Martin Potthast for their encouragement and kind support.

References

1. 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)
2. Cai, C., Li, L., Zengi, D.: Behavior enhanced deep bot detection in social media. In: 2017 IEEE International Conference on Intelligence and Security Informatics (ISI). pp. 128–130. IEEE (2017)
3. Clark, E.M., Williams, J.R., Jones, C.A., Galbraith, R.A., Danforth, C.M., Dodds, P.S.: Sifting robotic from organic text: a natural language approach for detecting automation on twitter. Journal of Computational Science 16, 1–7 (2016)
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. Davis, C.A., Varol, O., Ferrara, E., Flammini, A., Menczer, F.: Botornot: A system to evaluate social bots. In: Proceedings of the 25th International Conference Companion on World Wide Web. pp. 273–274. International World Wide Web Conferences Steering Committee (2016)
6. Dickerson, J.P., Kagan, V., Subrahmanian, V.: Using sentiment to detect bots on twitter: Are humans more opinionated than bots? In: Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. pp. 620–627. IEEE Press (2014)
7. Ferrara, E., Varol, O., Davis, C., Menczer, F., Flammini, A.: The rise of social bots. Communications of the ACM 59(7), 96–104 (2016)
8. Krzywicki, A., Wobcke, W., Bain, M., Martinez, J.C., Compton, P.: Data mining for building knowledge bases: techniques, architectures and applications. The Knowledge Engineering Review 31(2), 97–123 (2016)
9. 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)
10. 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)

11. 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)
12. Ratkiewicz, J., Conover, M.D., Meiss, M., Gonçalves, B., Flammini, A., Menczer, F.M.: Detecting and tracking political abuse in social media. In: Fifth international AAAI conference on weblogs and social media (2011)
13. 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)
14. Varol, O., Ferrara, E., Menczer, F., Flammini, A.: Early detection of promoted campaigns on social media. EPJ Data Science 6(1), 13 (2017)
15. Yang, K.C., Varol, O., Davis, C.A., Ferrara, E., Flammini, A., Menczer, F.: Arming the public with ai to counter social bots. arXiv preprint arXiv:1901.00912 (2019)