PAN @ CLEF 2019 - CELEBRITY PROFILING

BOTS AND GENDER PROFILING ON TWITTER USING SOCIOLINGUISTIC FEATURES

Edwin Puertas, Luis Gabriel Moreno Sandoval, Flor Miriam Plaza del Arco, Alexandra Pomares Quimbaya, Jorge Andrés Alvarado Valencia, and L. Alfonso Ureña López
Pontificia Universidad Javeriana, Bogotá, Colombia
{edwin.puertas,jorge.alavarado,morenoluis,pomares}@javeriana.edu.co
Center of Excellence and Appropriation in Big Data and Data Analytics (CAOBA)
Universidad de Jaén, Jaén, Andalucía, Spain.
{fmplaza, laurena}@ujaen.es

Summary

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.

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.

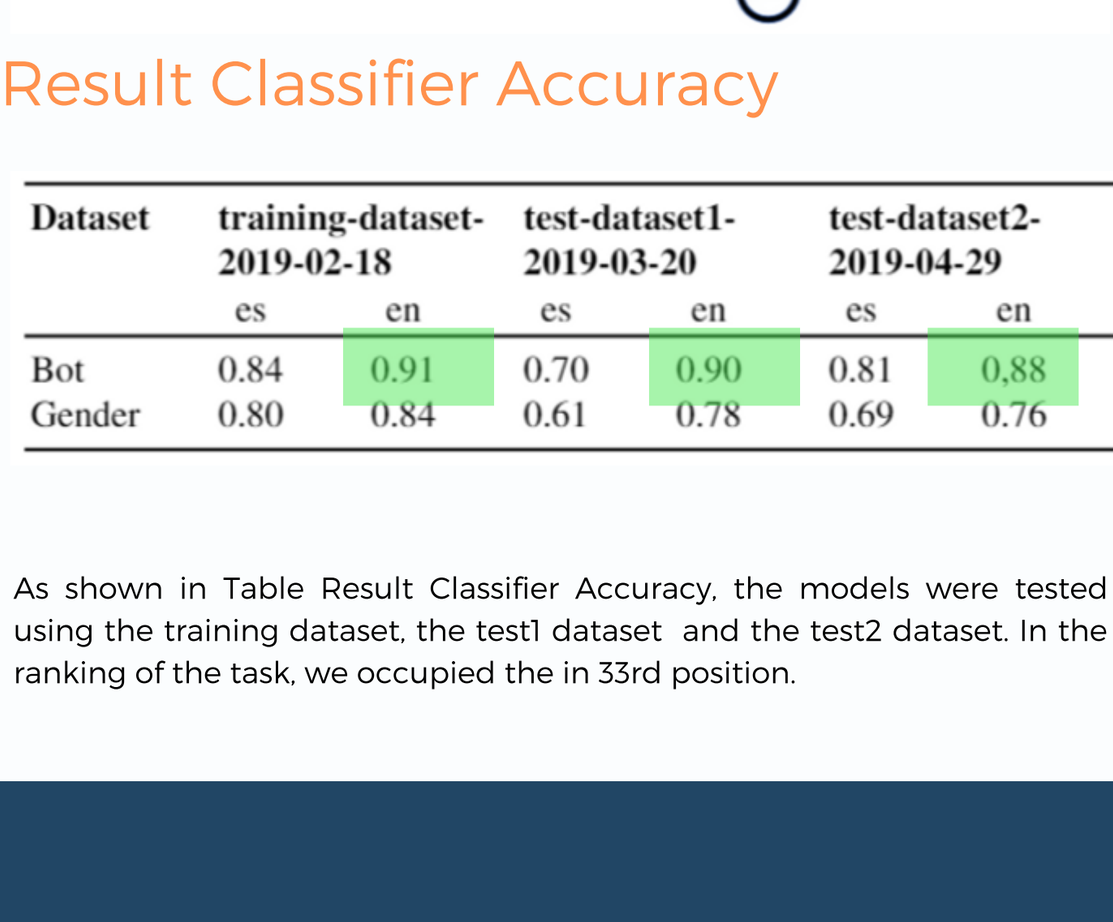
Features for Classification Model

#	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

+ Words Vector from Tweets

1) Network Features,
2) Lexical Features,
3) Sociolinguistic features
4) Text Tweet

Methodology System Training



Results

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.

Result Classifier Accuracy

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

As shown in Table Result Classifier Accuracy, the models were tested using the training dataset, the test1 dataset and the test2 dataset. In the ranking of the task, we occupied the in 33rd position.

ACKNOLEDGMENTS

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 construction of the research capabilities of CAOBA. Also, Fondo Europeo de Desarrollo Regional (FEDER), REDES project (TIN2015-65136-C2-1-R) and LIVING-LANG project (RTI2018-094653-B-C21) from the Spanish Government.