Bot and Gender Identification: Textual Analysis of Tweets

Notebook for PAN at CLEF 2019

Rodrigo Ribeiro Oliveira, Cláudio Moisés Valiense de Andrade, José Solenir Lima Figuerêdo, João B. Rocha-Junior, Rodrigo Tripodi Calumby, Iago Machado da Conceição Silva, Almir Moreira da Silva Neto

University of Feira de Santana
rodrigo18br@hotmail.com
{claudiovaliense, solenir.figueredo, iagomachado09, almirneto338}@gmail.com
{joao, rtcalumby}@uefs.br
Feira de Santana, Bahia, Brazil

Abstract In this paper, we describe the participation of the Advanced Data Analysis and Management (ADAM) group of the University of Feira de Santana in the *Bots and Gender Profiling Task* organized by PAN@CLEF 2019. We used Support Vector Machines (SVM) optimized through nested cross-validation. In bot detection, we used features related to behavior of the account, sentiment and variety of posts, in gender detection function words and emoticons. These features were evaluated both individually and in groups. Before starting the training phase, we preprocessed the data to better adjust it. For bot detection, our method reached approximately 0.9057 for English and 0.8767 for Spanish. For gender detection, 0.7696 for English and 0.7150 for Spanish. Although the results for Spanish are poorer than the ones for English, they are above the random baseline (50%).

1 Introduction

Social media companies employ mobile and web-based technologies to create highly interactive platforms through which individuals and communities share, cocreate, discuss, and modify user-generated content [10]. These services have changed the way we see the world and how information is disseminated. A prominent example of such services is Twitter[1]. Twitter is a popular microblogging service. Microblogging is a form of communication in which users must describe their current status in short posts distributed by instant messages, mobile phones, email or the Web [8]. In this kind of social media, users follow others or are followed. However, unlike other social networks, Twitter demands no reciprocity in the following-followed binomial. Thus, when following

[1] http://www.twitter.com

a particular user, that user is not required to follow you back. In this kind of application, users write about different aspects of their life, sharing a variety of subjects, and generating heterogeneous discussion.

Twitter is used in multiple contexts. It is mainly considered as an information dissemination tool, but also as a source of data that may support studies in different areas of knowledge. This feature is specially interesting considering it offers an Application Programming Interface (API)[2] that allows crawling and collecting data. In this context, a subject that has attracted the attention of researchers is the so-called Author profiling, in which information like age, cultural background, gender, native language, and personality can be inferred through textual analysis of users' posts. This type of analysis enables numerous applications, such as business intelligence, digital forensics, psychological profiling, brand reputation monitoring, etc. With regard to forensic applications, bot detection has gained attention, especially due to bots' self-controlled ability to disseminate political, extremist or misinformation material that may negatively influence a massive amount of users [6].

In this context, this paper describes the participation of the ADAM team in the Bots and Gender Profiling Task [18] organized by PAN@CLEF 2019. In this edition, different from previous years, in which various aspects of the author's profile in social media (age and gender, also along with personality, gender and variety of languages, and gender from a multimodal perspective) were investigated, this edition included, in addition, investigation whether the author of a Twitter feed is a bot or a human. The gender profiling was maintained as a task. The analysis, as in other editions, followed a multilingual perspective, with English and Spanish being the chosen languages. The main contributions of this paper are:

- We define a set of features with discrimative power for bot and gender detection;
- We evaluate the power of optimizing a model through cross-validation;
- We analyze the effectiveness of each group of features in the task.

The remainder of this paper is organized as follows. Section 2 presents the related work. Section 3 presents the experimental validation setup. The results are discussed in Section 4. Finally, Section 5 presents our conclusions and directions for future work.

2 Related Work

The use of bots has brought problems in many collaborative systems, e.g., Wikipedia and OpenStreetMaps. This scenario motivates the study of strategies to identify bots in such collaborative systems, examining the contributions done by users [7].

For textual classification, information about function words, n-grams (at word and character level), quantitative features, orthographic features, part of speech (POS) tags, and vocabulary richness features are usually used [11]. In a given country, the texts produced in a language may vary depending on the culture of the region of origin [15]. Language variety identification is a popular research topic of natural language processing [17]. The regional influence has an impact on the extraction of features, for example,

[2] https://developer.twitter.com/en/docs.html

in sentiment analysis, the weight of the dictionary used may vary according to location of the author.

In the gender identification task, previous works show that women use question tags more frequently, more emoticons, and less profanities [12]. The work in [19] identifies messages from men as often related to themes such as money, sports, and work, while women refer more frequently about family, friends, and food. In addition, some subjects are predominantly approached by men (e.g. gaming) and women (e.g. shopping).

The work in [1] explored some techniques for identifying twitter messages produced by bots. In the experiment, results are presented concerning the use of neural networks (MLP) and Random Forest. The Random Forest algorithm presented superior performance, reaching an accuracy of 92%. In comparison to the dataset of this article, the dataset of Braz el al. [1] contains additional information about the user accounts (e.g. number of users the account follows, number of followers), impacting the results achieved, because there is data beyong the textual aspect.

In [2], a purely textual approach to bot detection is used. Three features are build from the text: dissimilarity between pairs of tweets of a user; word introduction decay rate (a measure of new unique words a user introduced over time) and average number of URLs per tweet. The classification used 10-fold cross validation and achieved a rate of 90.32% in detecting bots.

Similar research is related to gender identification in e-mail messages [3]. Some features used are those that express emotions through a sequence of characters (e.g. “ur ssoooo kooool”, “ihaaa”) and emoticons (e.g. “:D”, “:(”). Another feature used by the authors in gender identification is the number of words ending with a sequence of characters (e.g. ‘less’). The paper suggests through various research that women make frequent use of adverbs and emotionally intensive adjectives and terms related to questions, personal orientation and support.

The work in [9] compares several classifiers and states that SVM is appropriate for classifying text, some of the reasons being the high dimensionality in the input set (large amount of features), and having overfitting protection.

3 Experimental Setup

In this section, we present the dataset (Section 3.1), the preprossessing phase (Section 3.2), the feature extraction method (Section 3.3) and the classification model (Section 3.4).

3.1 Dataset Description

Table 1 contains a description of the training corpus. The data was previously split in train and dev for both English and Spanish, indicating datasets for training and validation. All datasets were balanced between bot and human. Only the human portion of the data was labeled regarding gender, with equal number of male and female users also.

Each file corresponds to a user, containing 100 tweets from them. The tweets are not processed in any way, so emojis, retweets, hashtags, user mentions and hyperlinks are still present.

Table 1. Summary of dataset: number of files.

Language	Train	Dev
English	2880	1240
Spanish	2080	920

3.2 Preprocessing

In order to better adjust the posts to the experiments, we executed three independent operations:

- Conversion from multiple to single whitespaces;
- Lower-casing of all text;
- Removal of non-alphanumeric characters.

All the operations were performed using the Natural Language Toolkit (NLTK) [13].

3.3 Feature Extraction

Multiple classes of features were used for bot and gender detection. The following sections describe each one individually, according to the associated task.

Bot detection

Many of the features used in previous work to detect bots in twitter, as usernames, geodata, tweet intervals and following data, is unavailable in this case. Therefore, the features are limited to the twitter text.

- **Twitter features:** these features are related to the twitter profile of the users. Frequency of hashtags (#), frequency of mentions of users (@), frequency of retweets (occurrence of the string "rt") and frequency of links (occurrence of "http"). The rationale behind the first two is that bots tend to try to increase their reach inserting trending hashtags in their posts or mentioning multiple users to call their attention. Bots use to retweet content as a way to easily build a profile, and constant posting of links is typical behavior of spam bots.
- **Sentiment features:** in the English corpus, the VADER library[3] was used to identify the sentiments of the corpus. The VADER generates the mean of the four sentiment metrics: positive, negative, neutral and compound. An additional feature was also computed, the sentiment flip, defined in [5] as the number of sentiment inversions (positive to negative and vice-versa) between two adjacent posts normalized by the total number of tweets authored by the user. In VADER, the compound value has the range [-1, 1], so is the one used to calculate the sentiment flip, the point of inversion being 0. Getting the sentiment features for the Spanish corpus was somewhat difficult. The method recommended by the VADER developers is to translate

[3] https://github.com/cjhutto/vaderSentiment

the texts to English automatically and use the library on the resulting text. We ended using a machine learning based solution[4], which generated only one output in the range [0, 1]. The point of sentiment inversion was fixed on 0.5.

- **Variety features:** these features measure how varied is the content generated by the user, under the assumption that bots tend to repeat content in their posts. Two features belong to this group: the ratio between number of words used and total number of words in all posts; and the cleanliness: the ratio between the number of characters after and before preprocessing.

Gender detection

- **Emoticons:** frequency of each term in a list of emoticons, meant to describe a range of emotions.
- **Function words:** frequency of function words: pronouns, determiners, modals and conjunctions in English and in Spanish, conjunctions and determiners.
- **Sentiment features:** the same sentiment features used for bot detection.

3.4 Classification Model

For conducting the experiments we used a SVM classifier. In order to optimize the model learned, a 5-fold nested cross-validation [20] was done. In nested cross-validation, after a variation of parameters in order to find the optimal model, one more cross-validation is done to evaluate the model found.

This process is done only in the *train* dataset. After the model is chosen, it is used to classify the *dev* dataset, giving a notion of how the model will perform in unseen data, i.e., this set validates the model. To make this last step statistically significant, during the learning process nothing of the *dev* dataset was used.

In the cross-validation, both linear and RBF kernels were used varying their respecting hyperparameters (C and in RBF, *gamma*). To do this, we used the scikit-learn library [14].

4 Results and Discussion

The training and validation steps were done to each feature group individually at first. After that, the same was done using all the features together, so the impact of each feature group on the overall result could be evaluated.

The final evaluation was performed with the official PAN@CLEF 2019 [4] test set using the TIRA platform [16]. The results were grouped according to their associated task, i.e, bot or gender detection. For each task we present the assessment of the classifier for the three datasets provided: Train; Validation; and Test.

4.1 Bot detection

Table 2 presents the results of bot detection for the English portion of the data. In this experiment the overall best performance is achieved by using all the features, rather than using them separately.

[4] https://github.com/aylliote/senti-py

Table 2. Results for bot detection in English language.

Group	Train (%)	Validation (%)	Test (%)
Twitter Features	91.70	89.91	–
Sentiment Features	80.31	74.03	–
Variety Features	63.51	66.21	–
All features	93.12	90.97	90.57

Table 3 presents the results of bot detection for Spanish. The quite poor accuracy of the Sentiment-based features on train set, slightly above the random baseline (50%), led us to do the training experiments without this feature class, which had better results. Hence, we decided not to include this feature in the final submission for Spanish in both tasks.

Table 3. Results for bot detection in Spanish language.

Group	Train (%)	Validation (%)	Test (%)
Twitter Features	84.04	83.91	–
Sentiment Features	57.50	–	–
Variety Features	73.26	70.33	–
Twitter + Variety	86.16	87.07	87.67
All features	85.58	87.61	–

4.2 Gender Detection

Table 4 presents the results of gender detection for English. Similar to the Bot detection task, in the gender identification, the best performance occurs when using all features. Although the computational cost is higher, when using more features, it is expected that there will be an information gain, impacting positively the performance of the classifier.

Table 4. Results for gender detection in English language.

Group	Train (%)	Validation (%)	Test (%)
Function words	69.23	72.23	–
Function words + Sentiment Features	69.72	72.90	–
Function words + Sentiment Features + Emoticons	70.60	75.67	76.86

Table 5 presents the results in gender detection for Spanish. Adding emoticons to Function words improved the results only slightly, and the results were poorer than in English.

Table 5. Results for gender detection in Spanish language.

Group	**Train (%)**	**Validation (%)**	**Test (%)**
Function words	62.39	69.42	–
Function words + Emoticons	62.61	68.84	71.50

5 Conclusion and Future Works

In this paper we described the participation of the ADAM team in the Bots and Gender Profiling Task organized by PAN@CLEF 2019. In this task, focused on Twitter posts, should be determined if the author of a Twitter feed was a bot or human. Moreover, in case of a post from a human, the challenge was to identify the gender. We used a set of features in SVM with cross-validation.

The final outcome suggests that our proposal, in general, achieves good results when compared to the random baseline. The best results were achieved for Bot detection in English. Although the accuracy for gender detection is inferior to the accuracy in Bot detection, it also presents promising results.

As future work, we suggest trying to extract metadata from the tweet text, for example, building networks of citations through user mentions. Another approach is to detect how original are the posts of a user within the corpus, since bots are known to replicate human content to fake authenticity. In addition, it is interesting to experiment different machine learning approaches, such as deep learning. Better tools for features in non-English languages are needed, like the Sentiment ones, which are scarce.

References

1. Braz, P.A., Goldschmidt, R.R.: Redes neurais convolucionais na detecção de bots sociais: Um método baseado na clusterização de mensagens textuais. In: Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais (SBSeg), SBSeg 2018. pp. 323–336. SBC (2018)
2. 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)
3. Corney, M., De Vel, O., Anderson, A., Mohay, G.: Gender-preferential text mining of e-mail discourse. In: 18th Annual Computer Security Applications Conference (ACSAC), 2002. Proceedings. pp. 282–289. IEEE (2002)
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. 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)

6. Ferrara, E.: Disinformation and social bot operations in the run up to the 2017 french presidential election. First Monday 22(8) (2017)
7. Hall, A., Terveen, L., Halfaker, A.: Bot detection in wikidata using behavioral and other informal cues. Proceedings of the ACM on Human-Computer Interaction 2(CSCW), 64 (2018)
8. Java, A., Song, X., Finin, T., Tseng, B.: Why we twitter: Understanding microblogging usage and communities. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis. pp. 56–65. ACM, New York, NY, USA (2007)
9. Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. In: European conference on machine learning (ECML). pp. 137–142. Springer (1998)
10. Kietzmann, J.H., Hermkens, K., McCarthy, I.P., Silvestre, B.S.: Social media? get serious! understanding the functional building blocks of social media. Business Horizons 54(3), 241 – 251 (2011)
11. Koppel, M., Argamon, S., Shimoni, A.R.: Automatically categorizing written texts by author gender. Literary and linguistic computing (LLC) 17(4), 401–412 (2002)
12. Lakoff, R.: Language and woman's place. Language in society 2(1), 45–79 (1973)
13. Loper, E., Bird, S.: Nltk: the natural language toolkit. arXiv preprint cs/0205028 (2002)
14. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. Journal of machine learning research (JMLR) 12(Oct), 2825–2830 (2011)
15. Pennebaker, J.W., Mehl, M.R., Niederhoffer, K.G.: Psychological aspects of natural language use: Our words, our selves. Annual review of psychology 54(1), 547–577 (2003)
16. 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)
17. Rangel, F., Franco-Salvador, M., Rosso, P.: A low dimensionality representation for language variety identification. In: Gelbukh, A. (ed.) Computational Linguistics and Intelligent Text Processing. pp. 156–169. Springer International Publishing, Cham (2018)
18. 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)
19. 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)
20. Stone, M.: Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society: Series B (Methodological) 36(2), 111–133 (1974)