# Bots and Gender Profiling using Masking Techniques Notebook for PAN at CLEF 2019

Victor Jimenez-Villar<sup>1</sup>, Javier Sánchez-Junquera<sup>3</sup>, Manuel Montes-y-Gómez<sup>1</sup>, Luis Villaseñor-Pineda<sup>1,2</sup>, and Simone Paolo Ponzetto<sup>4</sup>

 <sup>1</sup> Instituto Nacional de Astrofísica Óptica y Electrónica, México. {victor.jimenez,mmontesg, villasen}@inaoep.mx
<sup>2</sup> Centre de Recherche en Linguistique Française GRAMMATICA (EA 4521), Université d'Artois, France.
<sup>3</sup> PRHLT Research Center, Universitat Politècnica de València, Spain. jjsjunquera@gmail.com

<sup>4</sup> Data and Web Science Group, University of Mannheim, Germany. simone@informatik.uni-mannheim.de

**Abstract** This work describes our proposed solution for the author profiling shared task at PAN 2019. The task consists in identifying whether the author of a Twitter feed is a bot or a human, and, in case of a human, in determining if the author is male or female. Like previous years, the task considers different languages, in this case, English and Spanish. Our proposal focuses on the preprocessing and feature extraction steps; we mainly apply some masking techniques that allow emphasizing the relevant terms by obfuscating the irrelevant ones but keeping information about the structure of the texts. Using this approach we obtained accuracies of 0.92 and 0.81 in the Spanish test set for classifying bots/humans and males/females, respectively; similarly, we obtained accuracy values of 0.91 and 0.82 for the English dataset.

# 1 Introduction

The use of bots in social media, like Twitter or Facebook, has been constantly increasing. One of the main purposes of bots is to influence the opinion of people in a particular matter [22]; sometimes the use of bots might be considered unethical which is the case when they are used in a political context [9]. Furthermore, bots are commonly related to fake news spreading and polarisation. Therefore, to approach the identification of bots from an author profiling perspective is of high importance from the point of view of marketing, forensics, and security [17]. The author profiling task consists in identifying some specific traits like gender or age, given a set of documents written by him/her. In this work, we focus on the preprocessing step, by which we mask irrelevant terms while maintaining the relevant ones, this with the objective of conserving a combination of style, content words, and the structure of the documents.

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). CLEF 2019, 9-12 September 2019, Lugano, Switzerland.

The rest of the paper is organized as follows. Section 2 presents a brief overview of the author profiling and bots detection tasks. Section 3 describes the proposed method. Section 4 describes the performed experiments and discusses the obtained results. Finally, Section 5 presents our conclusions.

## 2 Related Work

There are two main kinds of methods for bots detection: methods that use predictive features of individual users, and methods that consider the network structure to identify a coordinate group of bots [9]. A detailed overview of these methods is presented in [6].

Particularly, the network-based approaches look for anomalous behaviour in network interactions [13,9,27]. They have the advantage of requiring less data about the users and also being able to simultaneously detect multiple bots. On the contrary, the approach based on predictive features only relies on textual information [2,7,1]. There are also some hybrid approaches, for example, the work presented in [5]. It employs a Random Forest classifier trained with examples of both humans and bots behaviours, based on the Texas A&M dataset [10] which contains 15,000 examples of each class and millions of tweets. This work uses six different kinds of features related to the network, user metadata, friends, temporal data, content, and sentiment. Nevertheless, all these features are not always available, and therefore, in our work, we only make use of textual features.

Regarding gender profiling, a lot of significant work has been done. For example, in the previous edition of PAN@CLEF [18], several different approaches were evaluated, from traditional methods to deep learning techniques. Traditional methods employed content features such as word n-grams [4], and style features such as character n-grams [3] and counts of stop words, emojis, punctuation marks, etc [14]. Deep Learning approaches used word embeddings [20] as well as character n-grams embeddings [12], with a combination of different neural network architectures such as RNNs [23], CNNs, and bi-LSTMs [25]. Despite the use of more advanced techniques, traditional methods remain as a good and competitive approach. In this work, we use a unified approach for the two profiling tasks. It is based on a text distortion method proposed by Stamatatos to mask topic-related information [21]. Basically, it allows transforming the input texts into a more topic-neutral form while maintaining the structure of documents associated with the personal style of the authors.

# **3** Proposed Method

The proposed method is supported on the hypothesis that the general structure of a document in conjunction with a small subset of their words (mostly comprising style features but including a few content terms) are enough elements to characterize the author of a Twitter feed. With the objective of better understanding the results of the proposed method, we treat the author profiling task as two binary classification problems. First, we identify if the given author is a human or a bot. Then, in the case of a human, we determine if the author is male or female. Our method follows a classical

text classification pipeline considering preprocessing, feature extraction and classification. In this work, we put special emphasis on the preprocessing step. In the following subsections, the proposed method is described.

#### 3.1 Preprocessing

This phase is the most relevant in our method, it consists of two main steps: normalization and masking.

In the normalizing step, with the help of the python library Tweet Tokenizer from NLTK [11], we perform the following actions:

- 1. Join all tweets from each author in one string; tweets are separated by the END tag.
- 2. Replace line feed characters with NL.
- 3. Lowercase the characters.
- 4. Replace URLs with URL.
- 5. Replace user name mentions with USER.

It is important to mention that we maintain punctuation marks, emoticons and other special symbols.

The masking procedure is based on the work by Stamatatos [21]. Inspired by [8], he proposed a set of text distortion methods (DV-MA, DV-SA, DV-EX, DV-L2) in the context of authorship attribution. He found that the method DV-MA is more suited on texts related to blogs. The method DV-MA is defined as follows:

Let  $W_k$  be the list of k most frequent words of the language.

- Every word not included in  $W_k$  is masked by replacing each of its characters with an asterisk (\*).
- Every digit in the text is replaced by the symbol #.

For the task of authorship attribution masking all words except the k most frequent words is a good idea because it is mostly a style-based task, however, for the author profiling task it is important to consider style as well as some content features. Based on this observation, we consider different criteria to select the k most relevant terms:

**Document Frequency (DF)**: Defined as the number of documents in which the term  $t_i$  occurs. We compute the document frequency for each unique term in the corpus vocabulary, the assumption is that with this measure we can take a combination of style-related and topic-oriented terms.

**Frequently Co-occurring Entropy (FCE)**: This measure was proposed by [24] and used in [19] for domain adaptation deception detection. The objective of this measure is to pick out generalizable features that occur frequently in both classes and have similar occurring probability (for example, in both humans and bots).

**Information Gain (IG)**: It is commonly employed as a term selection criterion in classification tasks [26]. We compute the information gain of every term in the vocabulary and select as the k most relevant terms those with the greatest values, i.e., the subset of the most discriminative terms, which could include style and content elements.

#### 3.2 Feature Extraction

In order to find the best representation for the two author profiling tasks, we evaluated different representations based on word and character n-grams. Best results were obtained with 3-5 character n-grams. In all experiments we used this representation with the following parameters: we eliminated all n-grams occurring less than 3 times; we applied a  $tf \times idf$  weighting, with sub-linear term frequency (1 + log(tf)).

#### 3.3 Classification

In all the experiments we used a Support Vector Machine (SVM) classifier, with a linear kernel and default parameters according to the scikit-learn's implementation [15].

#### 4 Experiments and Results

As in previous years, the task's organizers provide a training corpus [17]. The corpus is composed of documents in English and Spanish, where each document contains 100 tweets for each author. The statistics of this corpus are presented in Table 1.

In order to compare the proposed method, we implemented three baselines: *i*) a traditional BOW model containing all words, *ii*) a BOW model applying a feature selection procedure, and *iii*) a character n-gram model applying the preprocessing actions but without masking.

Regarding the proposed method, we considered the different criteria to select the k most relevant terms: document frequency (DF), frequency co-occurring entropy (FCE), and information gain (IG). In all the cases, to find the optimal k value, we performed an incremental search using  $k = 100, 200, \dots 3000$ .

#### 4.1 Experimental Results

Table 2 shows the baseline results for the bots vs. human profiling task. We can observe that the model using character n-grams as features had the best results, an accuracy of 0.91 in Spanish and 0.93 in English. On the other hand, Table 3 shows the results of the proposed method (also with a representation of character n-grams) applying different criteria to select the k most relevant words (the words that were not masked). In this table we can observe that masking all words (i.e. using k = 0) produced a very competitive accuracy, 0.78 in Spanish and 0.83 in English, suggesting the importance of the text structure to distinguish bots from humans. In addition, we can observe that there are minor differences in the results from different selection criteria. Nonetheless, there is a slight improvement from 0.93 to 0.94 in English. Also, from the English results, we can deduce that the identification of bots and humans is entirely a stylistic task since the best results were obtained when selecting only the one hundred most frequent terms.

Tables 4 and 5 resume the results for the gender profiling task. For this task, the proposed method allowed improving the baseline results. In both languages, the best results were obtained when using the DF criterion. It is important to notice that for this

task it was necessary to maintain more terms unmasked (2600 and 2900 respectively), which indicates that, in contrast to the bots vs. human task, the differences between men and women are not only stylistic but also topic related. Furthermore, masking all the words did not produce good results, which means that texts generated by men and women share many structure characteristics.

Table 1. Number of documents by language.

| Author  | Training |         | Development |         |
|---------|----------|---------|-------------|---------|
|         | Spanish  | English | Spanish     | English |
| Males   | 520      | 720     | 230         | 310     |
| Females | 520      | 720     | 230         | 310     |
| Bots    | 1 040    | 1440    | 460         | 620     |

**Table 2.** Accuracy results of different baselines, in the classification of humans and bots. In the model BOW DF (Document Frequency) the best k value found was selected as a threshold, k=2900 for Spanish and k=100 for English. Classifying with SVM algorithm.

| Model              | Accuracy |         |  |
|--------------------|----------|---------|--|
| WIGUEI             | Spanish  | English |  |
| BOW                | 90       | 91.53   |  |
| BOW DF             | 89.89    | 90.48   |  |
| char n-grams (3-5) | 91.41    | 93.46   |  |

**Table 3.** Accuracy results applying the masking technique and using different feature selection methods, in the classification of humans and bots. The selected model was a combination of three to five character n-grams using a tf-idf weighting. Classifying with SVM algorithm.

| Selection | Spanish |          | English |          |
|-----------|---------|----------|---------|----------|
|           | k       | Accuracy | k       | Accuracy |
| None      | 0       | 77.93    | 0       | 82.98    |
| DF        | 2900    | 90.65    | 100     | 94.27    |
| FCE       | 2400    | 90.72    | 1000    | 93.7     |
| IG        | 1500    | 91.08    | 400     | 93.62    |

**Table 4.** Accuracy results of different baselines, in the classification of males and females. In the model BOW DF (Document Frequency) the best k value found was selected as a threshold, k=2600 for Spanish and k=2900 for English. Classifying with SVM algorithm.

| Model              | Accuracy |         |  |
|--------------------|----------|---------|--|
| WIUUCI             | Spanish  | English |  |
| BOW                | 64.13    | 81.61   |  |
| BOW DF             | 65.21    | 80.8    |  |
| char n-grams (3-5) | 66.3     | 80.96   |  |

**Table 5.** Accuracy results applying the masking technique and using different feature selection methods, in the classification of males and females. The selected model was a combination of three to five character n-grams using a tf-idf weighting. Classifying with SVM algorithm.

| Selection | Spanish |          | English |          |
|-----------|---------|----------|---------|----------|
|           | k       | Accuracy | k       | Accuracy |
| None      | 0       | 60.65    | 0       | 51.12    |
| DF        | 2600    | 69.13    | 2900    | 81.77    |
| FCE       | 2600    | 66.3     | 900     | 78.87    |
| IG        | 2900    | 66.3     | 2600    | 80.96    |

#### 4.2 Final Evaluation

In accordance with the previous results, for the final evaluation at TIRA platform [16], we applied our method selecting and maintaining the k terms with the greatest DF values, while the rest of them were masked. In particular, for the humans vs. bots profiling task we set k = 2900 for Spanish and k = 100 for English, whereas, for the gender profiling task we set k = 2600 for Spanish and k = 2900 for English.

The obtained accuracy results were as follows: in Spanish, 0.92 and 0.81 for classifying bots vs. humans and males vs. females, respectively; in English, 0.91 and 0.82 for both tasks. The results are very similar to those obtained in the training datasets, except for gender profiling in Spanish. We think this difference could be caused by the size of the training set on the final evaluation where the training and development partitions, used in the experimental configuration, were unified.

## 5 Conclusions

Distinguishing between bots and humans results to be not a very complex task. A traditional representation based on character n-grams showed to be a very strong baseline. Particularly our experiments showed that the single structure of the documents (i.e., when we masked all the words) provides important information for this classification task.

In the case of distinguishing between men and women, our experiments showed a very different picture. The proposed approach allowed improving the baseline results, suggesting that for this task it is important to have a combination of style and topic based features. the best results were obtained when maintaining the 2600-2900 more frequent terms and masking the rest.

As a general observation we can say that the proposed approach does not considerably improve the baseline results, but despite this, the masking technique has the advantage of normalizing the documents to a more neutral form and, therefore, of reducing the risk of overfitting.

Acknowledgments. This work was partially supported by CONACYT-Mexico under the scholarship 868585 and project grant CB-2015-01-257383.

## References

- Benevenuto, F., Magno, G., Rodrigues, T., Almeida, V.: Detecting spammers on twitter. In: Collaboration, electronic messaging, anti-abuse and spam conference (CEAS). vol. 6, p. 12 (2010)
- Chu, Z., Gianvecchio, S., Wang, H., Jajodia, S.: Detecting automation of Twitter accounts: Are you a human, bot, or cyborg? IEEE Transactions on Dependable and Secure Computing 9(6), 811–824 (2012)
- Daneshvar, S., Inkpen, D.: Gender identification in Twitter using N-grams and LSA: Notebook for PAN at CLEF 2018. In: Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). vol. 2125, pp. 1–10 (2018)
- von Daniken, P., Grubenmann, R., Cieliebak, M.: Word unigram weighing for author profiling at pan 2018. In: Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). vol. 2125. CEUR-WS.org (2018)
- 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)
- Ferrara, E., Varol, O., Davis, C., Menczer, F., Flammini, A.: The rise of social bots. Communications of the ACM 59(7), 96–104 (2016)
- Gianvecchio, S., Xie, M., Wu, Z., Wang, H., Member, S.: Humans and Bots in Internet Chat : Measurement, Analysis, and Automated Classification. IEEE/ACM Transactions on Networking 19(5), 1557–1571 (2011)
- Granados, A., Cebrian, M., Camacho, D., de Borja Rodriguez, F.: Reducing the loss of information through annealing text distortion. IEEE Transactions on Knowledge and Data Engineering 23(7), 1090–1102 (2010)
- 9. Hjouji, Z.e., Hunter, D.S., Mesnards, N.G.d., Zaman, T.: The impact of bots on opinions in social networks. arXiv preprint arXiv:1810.12398 (2018)
- Lee, K., Eoff, B.D., Caverlee, J.: Seven months with the devils: A long-term study of content polluters on twitter. In: Fifth International AAAI Conference on Weblogs and Social Media (2011)
- 11. Loper, E., Bird, S.: Nltk: the natural language toolkit. arXiv preprint cs/0205028 (2002)
- López-Santillán, R., Gonzalez-Gurrola, L., Ramfrez-Alonso, G.: Custom document embeddings via the centroids method: Gender classification in an author profiling task. In: Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). vol. 2125. CEUR-WS.org (2018)
- 13. Mesnards, N.G.d., Zaman, T.: Detecting influence campaigns in social networks using the ising model. arXiv preprint arXiv:1805.10244 (2018)
- Patra, B.G., Das, K.G.: Dd. multimodal author profiling for arabic, english, and spanish. In: Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). vol. 2125. CEUR-WS.org (2018)

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
- 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)
- 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)
- 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. CEUR Workshop Proceedings 2125 (2018)
- Sanchez Junquera, J.J.: Adaptación de Dominio para la Detección Automática de Textos Engañosos. Master's thesis, Instituto Nacional de Astrofísica, Óptica y Electrónica (2018)
- Schaetti, N.: Character-based convolutional neural network and resnet18 for twitter author profiling. In: Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). vol. 2125. CEUR-WS.org (2018)
- Stamatatos, E.: Masking topic-related information to enhance authorship attribution. Journal of the Association for Information Science and Technology 69(3), 461–473 (2018)
- Subrahmanian, V., Azaria, A., Durst, S., Kagan, V., Galstyan, A., Lerman, K., Zhu, L., Ferrara, E., Flammini, A., Menczer, F.: The darpa twitter bot challenge. Computer 49(6), 38–46 (2016)
- Takahashi, T., Tahara, T., Nagatani, K., Miura, Y., Taniguchi, T., Ohkuma, T.: Text and image synergy with feature cross technique for gender identification. In: Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). vol. 2125. CEUR-WS.org (2018)
- Tan, S., Cheng, X., Wang, Y., Xu, H.: Adapting naive bayes to domain adaptation for sentiment analysis. In: European Conference on Information Retrieval. pp. 337–349. Springer (2009)
- 25. Veenhoven, R., Snijders, S., van der Hall, D., van Noord, R.: Using translated data to improve deep learning author profiling models. In: Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). vol. 2125. CEUR-WS.org (2018)
- Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Icml. vol. 97, p. 35 (1997)
- Yang, Z., Wilson, C., Wang, X., Gao, T., Zhao, B.Y., Dai, Y.: Uncovering social network sybils in the wild. ACM Transactions on Knowledge Discovery from Data (TKDD) 8(1), 2 (2014)