Bots and Gender Identification Based on Stylometry of Tweet Minimal Structure and n-grams Model

Notebook for PAN at CLEF 2019

Alex I. Valencia Valencia[1], Helena Gomez Adorno[2], Christopher Stephens Rhodes[3], and Gibran Fuentes Pineda[2]

[1] Posgrado en Ciencia e Ingeniería de la Computación, UNAM, Mexico
letras_vivas@comunidad.unam.mx
[2] Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, UNAM, Mexico
helena.gomez@iimas.unam.mx, gibranfp@unam.mx
[3] Centro de Ciencias de la Complejidad, UNAM, Mexico
stephens@nucleares.com

Abstract Social media bots pose as humans to influence users with commercial, political or ideological purposes with the aim of artificially inflate the popularity of a product by promoting it or writing positive ratings and undermine the reputation of competitive products through negative valuations. The threat is even greater when the purpose is political or ideological. Therefore, to approach the identification of bots from an author profiling perspective is highly important for marketing, forensics, and security applications. For automatic bots automatic identification, we present an approach based on the tweet minimal structure and statistical metrics related to the entropy in every tweet. Using logistic regression, we achieved 86% and 90% of accuracy in the Spanish and English datasets respectively. In gender classification, we use an n-gram model including emojis and special characters converted to text data. We achieved 75% and 84% of accuracy in the Spanish and English datasets respectively, also using logistic regression classifier.

Keywords: Stylometry, Tweets, Bots Identification, Gender Profiling, n-grams, Logistic Regression, Entropy, Emojis, Special Characters

1 Introduction

Since origins, human communication has had big changes respecting data propagation, from word of mouth, radio, TV, etc. And now, society has been located in a digital era, called social media, a place where every user is a possible data propagator. However, this place is not populated only by humans but also by software-controlled agents, better known as bots. Bots are programmed to his creator intentions, from sending automated messages to assuming specific social or antisocial behaviors [13,12,7,14].

Similar to human interactions, bots can affect the structure and the function of a given society [2]. For this reason, detecting bots can help to maintain social stability, avoid network traps, and ensure the safety of privacy [3]. The Bots and gender profiling shared task at PAN 2019 [10], is focused on investigating whether the author of a Twitter feed is a bot or a human. Furthermore, in the case of human, to profile the gender of the author in English and Spanish languages.

The article is organized as follows. In Section 2, we show the state of the art related to bots and gender profiling. In Section 3, we briefly describe the Twitter corpus of the bots and gender profiling task at PAN 2019. In Section 4, we detail about minimal tweet structure approach used in this research, and we target the feature extraction, detailing on the preprocessing procedure for type and gender classification. Then, we show our results in Section 5. Finally, we rise our conclusions in Section 6.

2 Related Work

Due to social impact, there has been recent interest in bots automatic detection. The PAN 2019 [10] evaluation campaign included the author profiling task annually since 2013 and in this year included bots detection and gender profiling.

Kai-Cheng et al. [15] reviewed the literature on different bots, their impact, and detection methods. They used the case study of Botometer, a popular bot detection tool developed at Indiana University, to illustrate how people interact with AI countermeasures.

Stella et al. [12] analyzed a large-scale social data collected during the Catalan referendum for independence on October 1, 2017; consisting of nearly 4 million Twitter posts, and they identified two polarized groups of Independentists and Constitutionalists, quantifying the structural and emotional roles played by social bots.

Dickerson et al. [6] developed a collection of a network, linguistic, and application-oriented variables that could be used as features, and identify specific features that distinguish well between humans and bots. They analyzed a large dataset related to the 2014 Indian election, showing that a number of sentiment related factors are key to the identification of bots. The authors achieved a 0.73 value of area under the ROC Curve (AUROC).

Chiyu Cai et al. [3] proposed a behavior enhanced deep model (BeDM) for bot detection which regards user content as temporal text data instead of plain text to extract latent temporal patterns. BeDM fuses content information and behavior information using deep learning method achieving a 88.41 of precision value.

Concerning gender profiling, the second place at PAN 2017 used word unigrams and bigrams and character n-grams from 3 to 5 size as features moreover additional features include POS n-grams, emoji and document sentiment information [8].

More recently, at PAN 2018 [5] the best performing team in the gender profiling shared task retake the n-gram model. At character level, the authors used sizes of n-grams from 3 to 5. For the English dataset used at word level, unigrams, bigrams and trigrams. For the Spanish and Arabic datasets they used unigrams and bigrams of words. Furthermore, the winning approach used the LSA and TruncatedSVD function from the scikit-learn library.

3 Dataset Description

The training datasets for the author profiling task at PAN 2019 consists of tweets of human and bots authors, in case of human, can be male or female gender, both in English and Spanish languages. Comprises tweets of two groups of authors. In English there are 1440 bots, 720 Female and 720 Male. In Spanish, there are 1040 bots, 520 Female and 520 Male. The datasets are balanced on the type classification class, which is bot or human. For gender classification, the labels are bot, female or male. For each author (Twitter user), a total of 100 tweets were provided. Authors were coded with an alphanumeric author-ID.

4 Feature extraction and preprocessing

Depending on classification variable we used two different feature sets, in type classification we use statistical measures of different features and their corresponding entropy values. In the case of gender classification, we used the n-gram model proposed in [5].

4.1 Minimal tweet structure

The hypothesis is, that depending on bot design, normally these generate content automatically, thus leading to very similar tweet structure in each post. In this way, if we summarize the common elements from tweets these can contain. To this aim, we considered the following components: text, emojis, links, hashtags and user mentions. We can define a stylometry-based tweet minimal structure on the combination of these components.
For example, we have the following tweet: "I knew there was reason that @Kezzang69 went to work there.... https://t.co/vl6LgIHMMc"

We can identify the following tweet elements in post:

Text: I knew there was reason that
UserMention: @Kezzang69
Text: went to work there....
Link: https://t.co/vl6LgIHMMc

Given that the tweet is composed by a text section, then there is a user mention, followed by text and finally a link. Then, the minimal tweet structure for this post is:
['text', 'userMention', 'text', 'link']

4.2 Entropy of Minimal Tweet Structure

The next step is to measure the amount of information for every minimal tweet structure. We use the information entropy, which is associated with each data value and the negative logarithm of the probability mass function for the value:

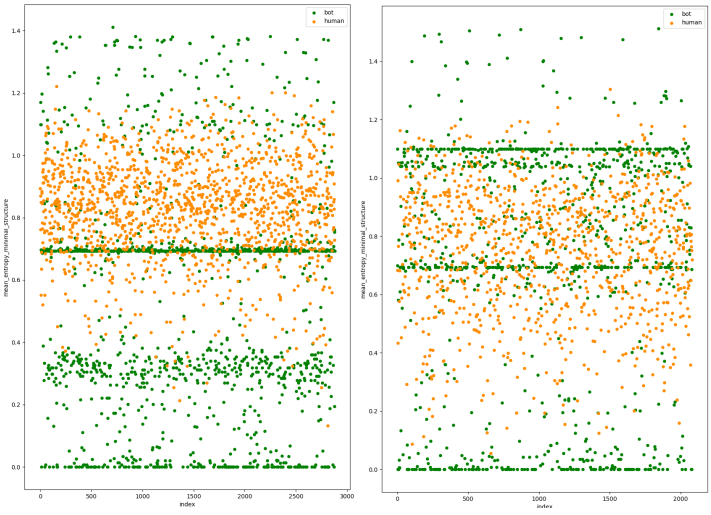
$$H(X) = \sum_{i=1}^{n} P(x_i) log_b P(x_i) \tag{1}$$

When the feature matrix produces a low probability value (i.e., when a low-probability event occurs), the event carries more "information" than when the source data produces a high-probability value [11]. In this way, following the hypothesis, bots should be associated with a low and/or a constant entropy value.

(a) English (b) Spanish

Figure 1. Entropy in the minimal tweet structure per user concatenating tweets

We compute the entropy of the training data set in two ways. In the first approach we concatenate the minimal structure of all tweets for every user and then compute entropy. Figure 1 shows the results for the English (a) and the Spanish (b) language.



(a) English (b) Spanish

Figure 2. Average entropy in the minimal tweet structure per user considering tweets individually

In the second approach we compute entropy for every tweet minimal tweet structure and then obtain the average of each entropy. Figure 2 shows the results of average entropy for English (a) and Spanish (b) languages.
In both figures, we can observe that the entropy of English bots are more distinctive from the entropy of humans than in the Spanish language, i.e. bots have a low and/or a constant entropy value, in English. It is important to mention that if we use only average entropy of minimal tweet structure (AEMTS) variable in a Gaussian Naive Bayes or Logistic Regression algorithms, we got 80% of accuracy.

4.3 Statistical Metrics

We also compute statistical metrics for each of the features mentioned above: text, emojis, links, hashtags, and user mentions. The following metrics were computed for each user: Sum, Maximum Value, Median, Average, Standard Deviation, Variance, and Entropy.

4.4 Preprocessing for Gender classification and Features

Based on [5] we added emojis and special characters rather than remove them; we use pandas and regular expressions to perform the next preprocessing steps:

1. Replace emojis with text using emoji library
2. Replace URLs with the word "link"
3. Replace User Mentions with the word "usermention"
4. Replace the linefeed characters with the word "linefeed"
5. Replace special characters with text using unicodedata library
6. Lowercased the characters
7. Trimmed the repeated characters: Replaced repeated character sequences of length 2 or grater with sequences of length 3
8. Any n-gram that occurred in all documents was considered a stop word and was ignored

5 Experimental settings

We examined the following algorithms for Type classification:

a) Naive Bayes
b) Logistic Regression with C=1e22
c) Support Vector Machine with linear kernel and C=1e6
d) Multi-Layer Perceptron with identity activation function, solver lbfgs, alpha 1e-5, 1000 maximum iterations and 200 in hidden layer sizes.

For Gender classification, we only experimented with b, c and d, we used the same parameters proposed in [5]. In all cases we use Normalizer[4] before the training process given that large margin classifiers are known to be sensitive to the way features are scaled [1].

[4] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.Normalizer.html

5.1 Gender Classification

In our experiments, during the development phase we got better results including emojis and special characters text data. In our final model we use the following n-grams features:

1. Words unigrams, bigrams and trigrams
2. Character 3-grams to 5-grams. In the same way that [5] we use TfidfVectorizer[5] with the following parameters for the feature sets:
   (a) Term frequency- inverse document frequency (tf-idf) weighting.
   (b) Sub-linear term frequency scaling, which uses $1 + log(TF)$ instead of TF.
   (c) Minimum document frequency = 0.01%: Terms with a document frequency strictly lower than 0.01% would be ignored.
   (d) Maximum document frequency = 1.0 (100%): Terms that occur in all documents would be ignored.

In the gender classification, we used only the n-grams model given that in our experiments they were better than AEMTS + metrics approach. In our experiments, we can see that adding emojis and special characters the performance was increased by 1%, as shown in Table 1.

Table 1. Gender Classification Results

Model / Test DataSet	**English-Dev**	**En-Official**	**Sp-Dev**	**Sp-Official**
n-grams	0.8496	N/A	0.7673	N/A
n-grams + Emojis and Special Characters	0.8626	0.8398	0.7728	0.7656

5.2 Type Classification

In the type classification, we used as features the Average Tweet Minimal Structure (AEMTS) as shown in Section 4, along with the statistical metrics. We evaluated our models on the official PAN 2019 [4] development and test sets for the author profiling task on the TIRA platform [9]. Moreover, we also evaluated the n-grams model in order to compare the accuracy achieved by both models, as shown in Table 2.

Table 2. Type Classification Results

Model / Test DataSet	**English-Dev**	**English-Official**	**Spanish-Dev**	**Spanish-Official**
AEMTS	0.82	N/A	0.78	N/A
AEMTS + Metrics	0.9444	0.9061	0.9082	0.8606
n-grams	0.921	N/A	0.8925	N/A
n-grams + Emojis and Special Characters	0.9354	0.9299	0.9112	0.9050

[5] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

As classification algorithm we trained a Logistic Regression classifier with the mentioned parameters. However, in the final evaluation on the test set, the performance decreased in comparison with the n-gram model. We believe that the proposed AEMTS + Metrics approach is a good and economic alternative to the n-gram model. The model trained on AEMTS + Metrics achieved competitive results on the development set.

6 Conclusions

For the author profiling task at PAN 2019 which comprises classifying two authors profiling aspects we can draw the following conclusions:

6.1 Type classification

Using Minimal tweet structure and entropy approach, have the following advantages: a) Variables are highly predictive for this data sets in bots detection b) Complexity order is lower than n-grams model generation c) Variables are independent of the language, i.e. no matter about words used by bots. Finally, hypothesis respect to bot behavior phenomenology was fulfilled correctly. In this way, depending on bots creators intention how much can change the tweet structure to be detected. The style frequency to differentiate from a human tweet entropy described in Section 5.

6.2 Gender classification

In the literature, generally the special characters and emojis are removed, however, we think all data can be information knowing how to deal it. In this case, adding the emojis and special character converted them to text data in our experiments improves the accuracy by 1%. On the other side, we think the selection of stop words is another key to get better performance.

References

1. Ben-Hur, A., Weston, J.: A user's guide to support vector machines. In: Data mining techniques for the life sciences, pp. 223–239. Springer (2010)
2. Bessi, A., Ferrara, E.: Social bots distort the 2016 u.s. presidential election online discussion. First Monday (November 2016)
3. 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 (July 2017)
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. In: Proceedings of the 9th International Conference of the CLEF Association (CLEF 2018). vol. 2125 (2018)
6. Dickerson, J.P., Kagan, V., Subrahmanian, V.S.: 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. ASONAM '14, IEEE Press, Piscataway, NJ, USA (2014), http://dl.acm.org/citation.cfm?id=3191835.3191957
7. 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
8. Martinc, M., Škrjanec, I., Zupan, K., Pollak, S.: Pan 2017: Author profiling-gender and language variety prediction (notebook for pan at clef 2017, 2nd place). In: CLEF (Working Notes). vol. 1866 (02 2018)
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., 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)
11. Shannon, C.E.: A mathematical theory of communication. Bell system technical journal 27(3), 379–423 (1948)
12. Stella, M., Ferrara, E., De Domenico, M.: Bots increase exposure to negative and inflammatory content in online social systems. Proceedings of the National Academy of Sciences 115(49), 12435–12440 (2018)
13. Tomasello, M.: Origins of human communication. MIT Press, Cambridge, Mass.; London (2010)
14. 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)
15. Yang, K., Varol, O., Davis, C.A., Ferrara, E., Flammini, A., Menczer, F.: Arming the public with AI to counter social bots. CoRR abs/1901.00912 (2019), http://arxiv.org/abs/1901.00912