Bots and Gender Profiling on Twitter
Notebook for PAN at CLEF 2019

Muhammad Hammad Fahim Siddiqui, Iqra Ameer, Alexander Gelbukh and Grigori Sidorov
Center for Computing Research (CIC),
Instituto Politécnico Nacional (IPN),
Av. Juan Dios Batiz, s/n, Zacatenco, 07738, Mexico City,
Mexico
{hammad.fahim57, iqraameer133}@gmail.com, {sidorov, gelbukh}@cic.ipn.mx

Abstract. This article describes the contribution of the Natural Language Processing Lab of CIC-IPN, Mexico in task bots and gender profiling at PAN-19 evaluation lab. The escalation in the use of social media facilities and proliferation in the fame of online social media websites such as Twitter, Facebook, LinkedIn, etc. directed to the growth of unwanted social bots as automatic social performers. Performers like this can perform numerous nasty acts comprising human discussions inflators, cheaters, stock market exploiters and so on. The risk is even higher when the motive is a political party. Moreover, bots are generally related to spreading fake news. So, it is essential to deal with the classification of social-bots from an author profiling point of view from the perspective of the marketing field, security, and forensics. We cannot deny the importance of bots on social media. Due to the high influence of bots, it is necessary to uncover the possible threats of social media bots. The goal of the article is to detect (A) if the author of a Tweet is a bot or a human, (B) if human, identify the gender of that particular author. We participated in the English language only. In the proposed approach, we used a well-known bag of words model with different preprocessing actions (stemming, stop words removal, lowercase, etc.). On provided development corpora, we got 87.12 (accuracy) on task A (binary class) by using Logistic Regression and 68.99 on task B (multi-class) by using Decision Tree classifier. In the evaluation phase on TIRA, we obtained 86.29 accuracy for task A and 68.37 for task B.

1 Introduction

An automated or social-bot is a program populating in social media websites that habitually interact with a human, trying to imitate and modify the behavior of individuals or creates fake content.

Bots on social media act like humans to affect users with a business, electoral or, political point of views. For instance, bots can work artificially to exaggerate the reputation of a product by writing positive ratings or/and supporting and can also damage the reputation of products in the competition via negative reviews. The hazard

is much higher when the intention is electoral. Be afraid of the impact of this effect; the political groups of Germany prohibited the utilization of bots in election movements for the general elections. Additionally, bots are connected to the spread of fake news; they can spread rumors [1]. In US media, it was broadly stated that public discussions impacted throughout the US 2016 presidential campaign because of false news reporting on social media. [2].

There can be good intentions behind the design of bots. Bots can perform good actions like updates of news and blog, which increase information distribution. They can be used to guard the privacy of members and perform duties faster than humans, such as asserting the news automatically, pushing the Wikipedia templates in a particular class to all pages and new updates [3], automatic distribution of a thank-you message to your new supporters. Bots can be useful such as simulated helpers for people like Siri or helping customers by providing a user-friendly facility [4] for the chatbots such as Microsoft's Tay and corporations, artificially intelligent bots. While conversely, bots can be proposed to perform nasty activities like (i) impact on stock markets: stock prices have been influenced by the bots. Investing choices are gradually being made by automatic trade systems that rapidly reply to the news on social media networks. (ii) Impact on economy: bots can damage the status of a firm or its products and lead to substantial economic loss. (iii) Cybercrime: researchers have determined how bots have got approach to private material, like cell numbers and addresses that can be utilized for cybercrime.

So, the classification of social bots from an author profiling perception is vital from the marketing perspective, security, and forensics.

The rest of the article is prepared as follows. In section 2, the existing work in the research community is described. In section 3, corpora provided by the PAN[1] [21] organizers and task description are presented. In chapter 4, details of our purposed approach and experiments to evaluate the system are described. In section 5, results and analysis are stated. Section 6 concludes the paper.

2 Related Work

Many investigators are working on the identification and profiling of bots in social media. To identify spam-bots, graph-based, and content-based features utilized, follow network connectivity and, extracted from the posted tweets respectively [5]. [6] examined if a Twitter account belongs to a bot, cyborg, or human. In this scenario, a bot was stated as spammy or a violent automatic profile, while cyborg denotes to a human-assisted bot or bot-assisted human. [6] also defined that, bots are more of malevolent type, and the research did not specify more investigation of malicious and benign bots in Twitter.

[7] formed honey-profiles on MySpace, Twitter, and Facebook to examine spambots. After a detailed analysis of the collected dataset, they recognized strange profiles who communicated the honey-profiles and created attributes for identifying spambots. Furthermore, 7-month research directed on Twitter by generating 60 social-

[1] https://pan.webis.de/clef19/pan19-web/author-profiling.html Last visited: 14/05/2019

honeypots that effort to trap spambots [8]. Twitter clients who message or follow more than two profiles of honeypot are instinctively supposed spambots. There is also research in the literature on spambot recognition due to social closeness [9] or social and content closeness [10]. It is defined in [11] who differentiated between bot accounts, managed accounts, and personal accounts of users on Twitter, according to time intervals of the tweet from the users.

In [12] developed a program to check if a Twitter profile acts similar to a bot or a human. They utilized the group of bots and human profiles distinguished by [8] and composed their tweets and track network information. In the 2014 Indian election, different features like linguistic, network, and application-oriented used to differentiate bots and humans [13]. [14] considered a network of bots for the study that mutually tweet concerning the war in Syria during 2012.

3 Corpora and Task Description

Bots and gender profiling on Twitter shared task at PAN 2019 had two datasets for the English and Spanish languages. However, we participated in the English language only.

3.1 Corpora

PAN-2019 provided us 412,000 labeled tweets of English language to train and develop the systems. Training corpus consisted of 288,000 labeled tweets, and 124,000 labeled tweets for the development phase (according to the PAN's suggested split of 70% for training and 30% for testing the models). The English training data set statistics are presented in Table 1, and statistics of development corpus are in Table 2. Different annotators manually labeled the corpora. More details can be found in overview papers [18, 22].

3.2 Description of the task

Task (A): if the author of a Tweet is a bot or a human: it is a binary classification problem, where it is asked to predict if a specific piece of text (tweet) written by a human or bot.

Task (B): if human, identify the gender of that particular author: It is multi-class classification problem, it is required to identify bot or human (e.g., the author of the specific tweet is human or bot) and in case of human, recognize the gender of human either male or female.

Table 1: English training corpus statistics.

Training corpus		
Human	**Human total**	**144000**
	Male	72000
	Female	72000
Bot		144000
Total instances		**288000**

Table 2: English development corpus statistics.

Development corpus		
Human	**Human total**	**62000**
	Male	15200
	Female	46800
Bot		62000
Total instances		**124000**

4 Description of our Approach

In this chapter, we defined our submitted approach considering the features and machine learning models used for this shared task.

4.1 Pre-processing

The given corpus was not cleaned by the organizers; they offered the tweets as they were tweeted by the users. Here the explanation is retweets were not taken out and multilingual tweets can appear. We applied preprocessing on the raw text before the extraction of features.

We performed the following steps:

- Performed stemming by using snow ball stemmer[2]
- Removed stop words
- Lowercased the text
- Punctuation marks were removed
- Removed HTML tags
- Converted the contracted forms into long forms, e.g., can't → cannot by using regular expressions
- Kept only alphabets, removed the numbers

[2] http://www.nltk.org/howto/stem.html Last visited: 14/05/2019

4.2 Features

The pre-processed text was used to generate the features for the machine learning (ML) algorithms. We used well-known TF-IDF values with n-gram range 1-3.

4.3 Machine learning algorithms

In our system, we tried a range of different classifiers for both tasks A and B, but we decided to mention best performing classifiers on our training dataset. For binary classification problem (Task A), we used Logistic Regression (LR), and for multi-class classification problem (Task B), we used Decision Tree (DT) classifier. For all classifiers, we used the available implementation in scikit-learn[3].

5 Results and Analysis

We are presenting our TIRA [19] results for both shared tasks, i.e., task A and B for the English language only. We used the following conventions. In the first column, the term "Tasks" refers to the shared tasks in which we participated. The name "Classifiers" states different classifiers, which we used in this competition. The term "Accuracy" refers to the evaluation measure used in this competition. To evaluate the systems, the PAN organizers calculated the accuracy of bot vs human. In case of humans, they calculated the accuracy of identifying males vs. females. At the end, they averaged the accuracy values per language to obtain the final ranking of the competition[4].

Table 3 is presenting the results on training corpora on TIRA platform. In the binary classification problem (human or bot), we got 97.94% accuracy by using LR classifier, which shows that the model is performing well on the binary classification task. On multi-class classification problem (in case of human, profile the gender), we achieved 77.01% accuracy by using DT algorithm.

Table 3: Results on training corpus.

Tasks	**Classifiers**	**Accuracy(%)**
Human/Bot (Task A)	Logistic Regression	97.94
Gender (Task B)	Decision Tree	77.01

Table 4: Results on development corpus-1.

Tasks	**Classifiers**	**Accuracy(%)**
Human/Bot (Task A)	Logistic Regression	87.12
Gender (Task B)	Decision Tree	68.99

Table 5: Results on development corpus-2.

Tasks	**Classifiers**	**Accuracy(%)**
Human/Bot (Task A)	Logistic Regression	86.29
Gender (Task B)	Decision Tree	66.29

[3] https://scikit-learn.org/ last visited: 18/05/2019

[4] https://pan.webis.de/clef19/pan19-web/author-profiling.html Last visited: 27/6/2019

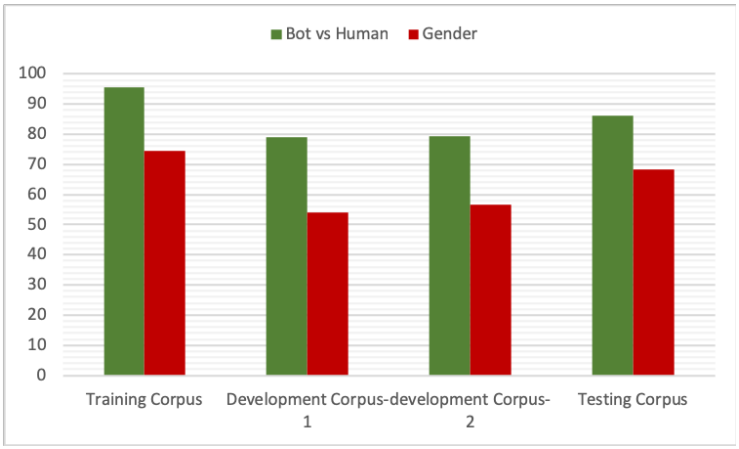

Figure 1:The trend of accuracies obtained for English language on training, development and testing corpora.

Table 4 is showing the results on development corpus-1, which is provided by the PAN-19 organizers to evaluate the model on TIRA settings. We got 87.12% and 68.99% accuracies on task A (binary) and task B (multi-class) respectively. Table 5 is providing the results on development corpus-2 (also provided by organizers) to evaluate the model. We acquired 86.29% and 66.29% accuracies on task A (binary) and task B (multi-class), respectively. In Figure 1, all results are reported including results in evaluation phase on TIRA. We achieved 86.29 accuracy for task A and 68.37 for task B. All reported results are for the English language.

6 Conclusion and Future Work

In the presented article, we explained our methodology to detect (A) if the author of a Tweet is a bot or a human, (B) if human, identify the gender of that particular author by using Twitter corpus. We participated in the English language only. We used TF and TF-IDF values with n-gram range 1-3. The vectors are then used as features for classifiers like LR and DT. Our model is performing well in the binary classification task by using development corpora provided by the organizers of PAN-19. Evaluation phase shows that the classification system is effective and correct to classify spambots and profile the gender on Twitter. In future, we can consider embeddings with TF-IDF weighting [15] and learning of document embeddings [16]. We also plan to work with syntactic n-grams (n-grams obtained by trailing paths in syntactic dependency trees) [17].

References

[1] C. Cai, L. Li, D. Zengi, "Behavior enhanced deep bot detection in social media," in Intelligence and Security Informatics (ISI), *2017 IEEE International Conference on. IEEE*, pp. 128–130 (2017).

[2] Williamson, W., Scrofani, J., Trends in Detection and Characterization of Propaganda Bots. In *Proceedings of the 52nd Hawaii International Conference on System Sciences,* Grand Wailea, Maui, (2019).

[3] (2016, September 12). Wikipedia: Creating a bot. Available: https://en.wikipedia.org/wiki/Wikipedia:Creating_a_bot.

[4] Freitas, C., Benevenuto, F., Ghosh, S., Veloso, A. Reverse engineering social bot infiltration strategies in twitter. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pp. 25-32, (2015).

[5] Wang, A. H. Detecting spam bots in online social networking websites: A machine learning approach. In *24th Annual IFIP WG 11.3 Working Conference on Data and Applications Security*, (2010).

[6] Chu, Z., Gianvecchio, S., Wang, H., Jajodia, S. Detecting automation of Twitter accounts: are you a human, bot, or cyborg?. *IEEE Trans. Dependable SecureComput*.9(6), 811–824, (2012).

[7] Stringhini, G., Kruegel, C. and Giovanni Vigna. Detecting spammers on social networks. In *Proceedings of the 26th Annual Computer Security Applications Conference*, ACM, 1–9, (2010).

[8] Lee, K., Eoff, B. D, Caverlee, J. Seven months with the devils: A long-term study of content polluters on Twitter. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 185–192, (2011).

[9] Ghosh, S., Viswanath, B., Kooti, F., Sharma, N. K, Korlam, G., Benevenuto, F., Ganguly, N., Gummadi, K. P. Understanding and combating link farming in the twitter social network. In *Proceedings of the 21st international conference on World Wide Web, WWW '12,* (2012).

[10] Hu, X., Tang, J., Zhang, Y., Liu, H. Social spammer detection in microblogging. In *Proceedings of IJCAI*, (2013).

[11] Tavares, Gabriela, Faisal, A.Scaling-Laws of Human Broadcast Communication Enable Distinction between Human, Corporate and Robot Twitter Users. *PLoS ONE* 8 (7): e65774, 2013.

[12] Ferrara, E., Varol, O., Davis, C., Menczer, F., Flammini, A. The rise of social bots. Comm. *ACM* 59(7):96–104, 2016.

[13] John, P., Dickerson, Kagan, V., Subrahmanian, V. S. Using Sentiment to Detect Bots on Twitter: Are Humans More Opinionated Than Bots*? Proc. IEEE/ACM Int'l Conf. Advances in Social Networks Analysis and Mining (ASONAM 14*), pp. 620–627, 2014.

[14] Abokhodair, N., Yoo, D., McDonald, D.W. Dissecting a social botnet: Growth, content, and influence in Twitter. In *Proceedings of the 18th ACM Conference on Computer-Supported Cooperative Work and Social Computing,* (2015).

[15] Arroyo-Fernández, I., C., Méndez-Cruz, Sierra, G., Torres-Moreno, J., Sidorov, G. Unsupervised Sentence Representations as Word Information Series: Revisiting TFIDF. *arXiv preprint arXiv*:1710.06524, (2017).
[16] Gómez-Adorno, H., Posadas-Durán, J., Sidorov, G., Pinto, D. Document embeddings learned on various types of n-grams for cross-topic authorship attribution. *Computing*. pp. 1-16, (2018).
[17] Sidorov, G. Syntactic N-grams in Computational Linguistics. Springer, 125 p, (2019).
[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] 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).
[20] Rangel, F., Rosso, P., Franco, M. A Low Dimensionality Representation for Language Variety Identification. In: Proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'16), Springer-Verlag, LNCS(9624), pp. 156-169, (2018).
[21] 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).
[22] Rangel, F., Rosso, P. Overview of the 7th Author Profiling Task at PAN 2019: Bots and Gender Profiling. In: Cappellato L., Ferro N., Müller H, Losada D. (Eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings. CEUR-WS.org