Bots and Gender Classification on Twitter

Notebook for PAN at CLEF 2019

Usman Saeed and Dr. Farid Shirazi
Data Science Lab (DSL),
Ryerson University, Canada,
350 Victoria St, Toronto, ON M5B 2K3, Canada.
{usman.saeed, f2shiraz}@ryerson.ca

Abstract. In the modern era, we observed a massive increase in the activities of social-media due to suitable large group of users. As Twitter is highly popular social networking site, Twitter has also appealed the interests of the spammers like social bots to behave as social media actors. Actors like this can perform many wicked actions, including individual discussion inflators, swindler, and stock market exploiters, and so on. The hazard is even higher when the purpose is a political party. Furthermore, bots are usually associated with spreading fake contents. So, it is vital to deal with the classification of bots from an author profiling point of view from the perception of the marketing field, network security, and Data forensics. This article describes the contribution of the Data Science Lab of Ryerson University, Canada in task bots and gender profiling at PAN-19 evaluation lab. The goal of this paper 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 bag of words model after applying different preprocessing techniques (stemming, stop words removal, lowercase, etc.). On the development dataset which was made available by PAN, we got best accuracies 79.51 on task A (binary class) by using MultinomialNB and 56.55 on task B (multi-class) by using Decision Tree classifier. In the evaluation phase on TIRA, our results are the same as in development dataset-2.

1 Introduction

A social-bot is an automated program that creates web content and attempts to interact with humans on social media platforms. Recently, we realized significant progress of actions by users presence in social-media platforms. For instance, Twitter advanced from a private micro-blogging platform to an information distribution platform. It became easy for a new user to set up an account due to possible access and openness of the Twitter platform. This set up allows the bot to post tweets like a human. There are both bad and good outcomes by the proliferation of bots [1, 2]. On one side, bots can produce some good, informative tweets like blog updates and news, which improve information broadcasting. Automated bots can also be useful for a profile holders, like bots that combined data from many information origins ground on the account holders' attentiveness. Contrarily, spammers and hackers can manipulate bots to appeal current

profiles as their supporters, allowing them to take over outcomes of the searching engines or running topics, distribute unwanted communications (messages, email, etc.), and tempt the users to visit malicious websites [2, 3, 4]. Furthermore, hackers can use malicious bots that can produce more severe effects like generating panic in emergencies, leaning political opinions, or harming a company's status [1, 5] and hack IT network. Therefore, this article uncovers the possible threats of nasty social bots, evaluations of the detection techniques and suggests possible paths for future study.
The rest of the article is arranged as follows. In section 2, the existing work in research community is explained. In section 3, dataset provided by the PAN[1] [16] 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 specified. Section 6 concludes the paper.

2 Related Work

Spam recognition examined for quite a while. The earlier work centers around spam-email recognition and identification of spam contents on web.

In [6], author first suggested a Bayesian method to clarify spam e-mails. Research outcomes display that the algorithms has an enhanced scores studying domain-specific features along with the unprocessed messages of E-mails. Presently spam e-mail cleaning is a moderately advanced method. Bayesian spam e-mail filters are executed both on modern e-mail users and servers. [7] formed honey-profiles on MySpace, Facebook and Twitter to examine spambots. After a full examination of the gathered dataset, they determined unexpected user profiles who connected the honey-profiles and formed attributes for classifying spambots. Additionally, research of seven months engaged on Twitter by producing 60 honeypots that try to trap spambots [8]. Twitter users who sent a message or followed two or more accounts of honeypot are instinctively supposed to be spambots. There is also a study in the research community on spambot identification grounded on social familiarity [9] or social and content familiarity [10]. It is described in [11] who distinguished among bot accounts, managed accounts, and personal accounts of clients on Twitter, based on time intervals of the tweet from the users.

In [12] developed an algorithm to check if a Twitter profile performs same as bot or an individual. They utilized the group of bots and individual profiles prominent by [8] and gathered their tweets and track network information. In 2014 Indian election, different features like linguistic, network, and application-oriented used to differentiate bots and individuals [13]. [14] considered a set-up of bots for the study that mutually tweet concerning the 2012 Syrian civil war.

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

3 Dataset and Task Description

The organizers of shared task bots and gender profiling on Twitter provided English and Spanish language datasets. However, we only participated in the English language.

3.1 Corpora

PAN-2019 released 412,000 labeled tweets of English language for the training and development of the systems. Dataset for the training of the model consists of 288,000 labeled tweets, and the development dataset includes 124,000 labeled tweets (according to the PAN's suggested split of 70% for training and 30% for development phase). The English training data set statistics are presented in Table 1 and statistics of development dataset are in Table 2. Various annotators manually labeled the dataset. Details can be found in overview paper [15].

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 task, where it is remanded to classify if a tweet written by a human or bot. The systems are ranked by accuracy.

Task (B): if human, identify the gender of that particular author: It is multi-class classification task, where asked to classify bot or human (e.g., the author of the specific tweet is human or bot) and in case of human, identify the gender of human either male or female. The systems are ranked by accuracy [15].

4 Description of our Approach

In this chapter, we describe our purposed approach considering the attributes and machine learning methods used for this shared task.

4.1 Pre-processing

The released dataset was not preprocessed; organizers provided the tweets as they were tweeted by the users. Here explanation is RTs were not removed and there are chances to appear multilingual tweets. Before the extraction of features, we applied preprocessing on raw text. Preprocessing helps to increase accuracy in classification tasks.
We performed the following steps:

- Removed stop words
- Lowercased the text

Table 1: English training dataset statistics.

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

Table 2: English development dataset statistics.

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

- Punctuation marks are removed
- Removed HTML tags
- Changed the contracted forms into long forms e.g. haven't → have not by using regular expressions
- Removed the numbers, kept only alphabets,
- Performed stemming by using snow ball stemmer[2]

4.2 Features

The cleaned text was used to generate the features for the machine learning (ML) algorithms. We used TF-IDF values with unigram bigram and trigram.

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.

Table 3: Results on training dataset.

Tasks	**Classifiers**	**Accuracy(%)**
Human/Bot (Task A)	MultinomialNB	95.73
Gender (Task B)	Decision Tree	74.34

Table 4: Results on development dataset-1.

Tasks	**Classifiers**	**Accuracy(%)**
Human/Bot (Task A)	MultinomialNB	79.17
Gender (Task B)	Decision Tree	54.17

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

Table 5: Results on development dataset-2.

Tasks	Classifiers	Accuracy(%)
Human/Bot (Task A)	MultinomialNB	79.51
Gender (Task B)	Decision Tree	56.55

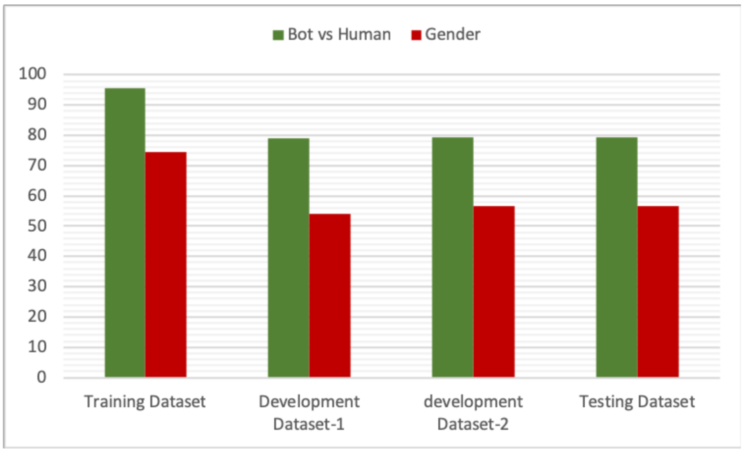

Figure 1:The trend of accuracies obtained for English language on training, development and testing corpora.

For binary classification problem (Task A), we used MultinomialNB (MNB), and for multi-class classification problem (Task B), we used Decision Tree (DT) classifier, For all classifiers, we used existing implementation in scikit-learn[3].

5 Results and Analysis

Shared results of TIRA[17] is presented , i.e., task A and B for the English language only. We used the following conventions. First column refers to the Shared task which we participated in. The second column “Classifiers” state different classifiers, which we used in this competition. Third column “Accuracy” points to the evaluation measure used in this competition.
Table **3** is presenting the results on training dataset on TIRA platform. On binary classification problem (Human or bot), we got 95.73% accuracy by using MNB 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 74.34% accuracy by using DT algorithm. Table **4** is showing the results on development dataset-

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

1, which is provided by the PAN-19 organizers to evaluate the model on TIRA settings. We got 79.17% and 54.17% accuracies on task A (binary) and task B (multi-class) respectively. Table **5** is providing the results on development dataset-2 (also provided by organizers) to evaluate the model. We acquired 79.51% and 56.55% 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. In evaluation phase, our results are same as on development dataset-2. 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] Chu, Z., Gianvecchio, S., Wang, H., Jajodia, S. Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE Tran Dependable & Secure Comput* 9(6):811–824, (2012).

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

[3] 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).

[4] Hu, X., Tang, J., Zhang, Y., Liu, H. Social spammer detection in microblogging. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, (2013).

[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] Sahami, M., Dumais, S. David Heckerman, and Eric Horvitz. A bayesian approach to filtering junk e-mail. In *AAAI-98 Workshop on Learning for Text Categorization*, (1998).

[7] Stringhini, G., Kruegel, C., Vigna, G. 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, Vadim Kagan, and V.S. Subrahmanian. 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. and 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] 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, (2019).

[16 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).

[17] 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).

[18] 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.

[19] 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