Twitter User Profiling: Bot and Gender Identification

Notebook for PAN at CLEF 2019

Dijana Kosmajac and Vlado Keselj

Dalhousie University
dijana.kosmajac@dal.ca, vlado@dnlp.ca

Abstract We use a set of feature extraction and transformation methods in conjunction with ensemble classifiers for the PAN Author Profiling task. For the bot identification subtask we use user behaviour fingerprint and statistical diversity measures, while for the gender identification subtask we use a set of text statistics, as well as syntactic information and raw words.

1 Introduction

Automated user (bot) is a program that mimics a real person's behavior on social media. A bot can operate based on a simple set of behavioral instructions, such as tweeting, retweeting, "liking" posts, or following other users. In general, there are two types of bots based on their purpose: non-malicious and malicious. The non-malicious bots are transparent, with no intent of mimicking real Twitter users. Often, they share motivational quotes or images, tweet news headlines and other useful information, or help companies to respond to users. On the other hand, malicious ones may generate spam, try to access private account information, trick users into following them or subscribing to scams, suppress or enhance political opinions, create trending hashtags for financial gain, support political candidates during elections [2], or create offensive material to troll users. Additionally, some influencers may use bots to boost their audience size.

We explore bot and gender identification techniques on PAN 2019 [5] Author Profiling task [19]. We apply a set of feature extraction methods to describe how diverse the user behaviour is over extended period of time and if the style of writing is different between two genders. The systems were hosted and evaluated on TIRA [18], a web service that aims to facilitate software submissions and evaluations for shared tasks.

The rest of the paper is organized as follows. Related work is discussed in Section 2. Section 3 briefly shows insights into the datasets. Section 4.1 describes the method we used to extract and encode features in the form of digital fingerprint. In Section 4 we describe a set of features used for user profiling, for both gender and bot identification tasks. Section 5 is dedicated to experiments and results. Finally, in Section 6 we give the conclusions and briefly discuss about future work.

2 Related Work

One of the most prominent tasks in recent social media analysis is detection of automated user accounts (bots). Research on this topic is very active [16,28,10], because bots pose a big threat if they're intentionally steered to target important events across the globe, such as political elections [2,27,14,12,23,13]. Paper by [16] explore strategies how bot can interact with real users to increase their influence. They show that a simple strategy can trick influence scoring systems. BotOrNot [6] is openly accessible solution available as API for the machine learning system for bot detection. Authors [6,27] show that the system is accurate in detecting social bots. Authors [21] explore methods for fake news detection on social media, which is closely related to the problem of automated accounts. They state that the performance of detecting fake news only from content in general doesn't show good results, and they suggest to use user social interactions as auxiliary information to improve the detection. Ferrara et al. [8] use extensive set of features (tweet timing, tweet interaction network, content, language, sentiment) to detect the online campaigning as early as possible. Another recent work on bot detection by Cresci et al. [3] is based on DNA inspired fingerprinting of temporal user behaviour. They define a vocabulary B^n, where n is the dimension. An element represents a label for a tweet. User activity is represented as a sequence of tweets labels. They found that bots share longer common substrings (LCSs) than regular users. The point where LCS has the biggest difference is used as a cut-off value to separate bots from genuine users. Framework by Ahmed et al. [1] for bot detection uses the Euclidean distance between feature vectors to build a similarity graph of the accounts. After the graph is built, they perform clustering and community detection algorithms to identify groups of similar accounts in the graph.

Bot problem on social media platforms inspired many competitions and evaluation campaigns such as DARPA [24] and PAN[1].

When it comes to gender and age user profiling, advances in natural language processing technology have facilitated the prediction in several text genres using automatic analysis of the variation of linguistic characteristics. However, in social media texts, there are a couple of limitations. First, small amount of meta information about the users' gender, age, social class, race, geographical location, etc., is available to researchers. Second, communication in online social networks typically occurs in a form of very short messages, often containing non-standard language usage, which makes this type of text a challenging text genre for natural language processing. Finally, given the speed at which chat language has originated globally and continues to develop, especially among young people, a third challenge in automatically detecting false profiles on social networks will be the constant retraining of the machine learning algorithms in order to learn new variations of chat language. Many researchers have tried to solve some of these challenges [20,17,11,25,4,17].

[1] https://pan.webis.de/publications.html

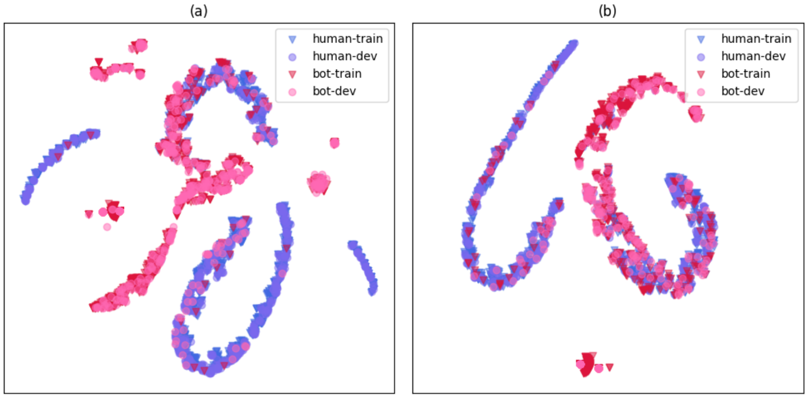

Figure 1. Bot t-SNE visualization. (a) English, (b) Spanish

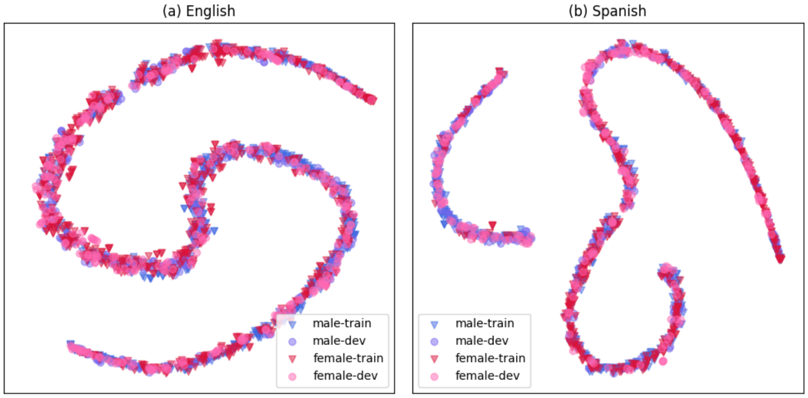

Figure 2. Gender t-SNE visualization. (a) English, (b) Spanish

3 Dataset

The dataset provided by the organizers is divided into two parts: English and Spanish. The English dataset consists of training and development subsets, with 2,880 and 1,240 samples, respectively. The Spanish dataset is slightly smaller and consists of training and development subsets, with 2,080 and 920 samples, respectively. Each sample is a user timeline in chronological order, with 100 messages per user. Fig. 1 and Fig. 2 show the datasets using t-SNE [15], an enhanced method based on stochastic neighbour embedding. The features used for both visualizations are the ones used for the classifiers in the final submitted run (Experiment 2(4) for bots, and Experiment 5 for gender).

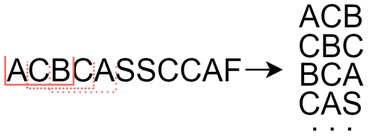

Figure 3. 3-gram extraction example from user fingerprint.

4 Feature Engineering

In this section we describe the features used for the experiments.

4.1 Bot Identification

User Behaviour Fingerprint DNA sequences have been exploited in different areas such as forensics, anthropology, bio-medical science and similar. Cresci [3] used the idea of DNA coding to describe social media user behaviour in temporal dimension. The same idea was used in this study, with a slightly modified way of coding. We define a set of codes A_n with length $n = 6$. The meaning of each code is given in (1).

$$A_n = \begin{cases} 0, & \text{plain} \\ 8, & \text{retweet} \\ 16, & \text{reply} \\ 1, & \text{has hastags} \\ 2, & \text{has mentions} \\ 4, & \text{has URLs} \end{cases} \tag{1}$$

Vocabulary, given the code set A, consists of $3 * 2^3 = 24$ unique characters. Each character, which describes a tweet is constructed by adding up codes for tweet features. First three codes describe the type of the tweet (retweet, reply, or plain) and the rest describe the content of the tweet. For example, if a tweet is neither retweet nor reply, it is plain (with the $code = 0$). If the tweet contains hashtags, then $code = code + 1$, If the same tweet contains URLs, then $code = code + 4$. Final tweet code is 5. We transform it to a character label by using ASCII table character indexes: $ASCII\_tbl[65+5] = F$. The number of tweets with attributes encoded with characters determines the length of the sequence. The sequence, in our case, is simply the length of a user timeline, that is, actions in chronological order with the appropriate character encoding.

The example of a user fingerprint generated from their timeline looks like:
$fp_{user} = (ACBCASSCCAFFADADFAFASCB...)$

Fingerprint segmentation using n-gram technique To calculate data statistics, we extracted n-grams of different length (1-3 length appeared to work best). Fig. 3 shows the example on 3-gram extraction of sample user fingerprint.

N-gram segments are used to calculate richness and diversity measures, which seem to unveil the difference between genuine user and bot online behaviour.

Statistical Measures for Text Richness and Diversity Statistical measures for diversity have long history and wide area of application [26]. A constancy measure for a natural language text is defined, in this article, as a computational measure that converges to a value for a certain amount of text and remains invariant for any larger size. Because such a measure exhibits the same value for any size of text larger than a certain amount, its value could be considered as a text characteristic. Common labels used are: N is the total number of words in a text, $V(N)$ is the number of distinct words, $V(m,N)$ is the number of words appearing m times in the text, and m_{max} is the largest frequency of a word.

Yule's K Index Yule's original intention for K use is for the author attribution task, assuming that it would differ for texts written by different authors.

$$K = C\frac{S_2 - S_1}{S_1^2} = C\Big[-\frac{1}{N} + \sum_{m=1}^{m_{max}} V(m,N)(\frac{m}{N})^2\Big]$$

To simplify, $S_1 = N = \sum_m V(m,N)$, and $S_2 = \sum_m m^2 V(m,N)$. C is a constant originally determined by Yule, and it is 10^4.

Shannon's H Index The Shannon's diversity index (H) is a measure that is commonly used to characterize species diversity in a community. Shannon‘s index accounts for both abundance and evenness of the species present. The proportion of species i relative to the total number of species (p_i) is calculated, and then multiplied by the natural logarithm of this proportion ($ln(p_i)$). The resulting product is summed across species, and multiplied by -1.

$$H = -\sum_{i=1}^{V(N)} p_i ln(p_i)$$

$V(N)$ is the number of distinct species.

Simpson's D Index Simpson's diversity index (D) is a mathematical measure that characterizes species diversity in a community. The proportion of species i relative to the total number of species (p_i) is calculated and squared. The squared proportions for all the species are summed, and the reciprocal is taken.

$$D = \frac{1}{\sum_{i=1}^{V(N)} p_i^2}$$

Honoré's R Statistic Honoré (1979) proposed a measure which assumes that the ratio of hapax legomena $(1,N)$ is constant with respect to the logarithm of the text size:

$$R = 100\frac{log(N)}{1 - \frac{V(1,N)}{V(N)}}$$

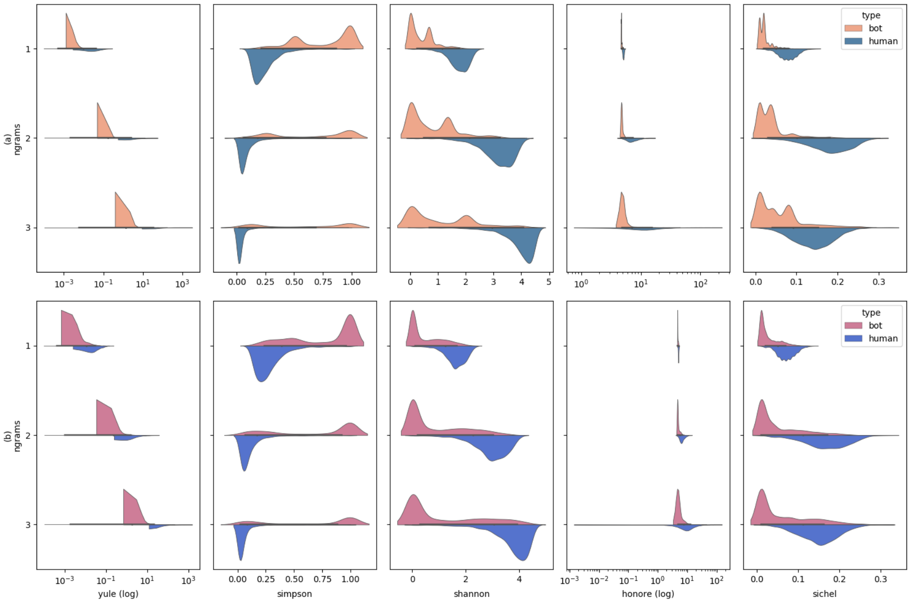

Figure 4. Diversity measures density per dataset, per user type. (a) English – top row, (b) Spanish – bottom row

Sichel's S Statistic Sichel [22] observed that the ratio of hapax dis legomena $V(2, N)$ to the vocabulary size is roughly constant across a wide range of sample sizes.

$$S = \frac{V(2, N)}{N}$$

We use this measure to express the constancy of n-gram *hapax dis legomena* (number of n-grams that occur two times) which we show to be distinct for genuine and bot accounts.

On Fig. 4 we show the comparison of density plots of all measures of bot accounts versus genuine users. We can see that the diversity measures are different for bots and genuine users. We exploit this characteristic to build a good classifier with as few features as possible.

4.2 Gender Identification

The feature types used for this task can be split into four categories:

Character and Word Features We used simple text metrics, such as total number of characters, total number of words, number of characters/words per message, number of special characters, number of digits.

PoS Tags Features Using spacy[2] python library we extracted word unigrams and bigrams, as well as PoS tag bigrams.

Emoji Features We counted the number of emojis, as well as fine-grained distinction between different types of emojis. To distinguish categories of emojis we used the latest standard at the time of experiments[3].

Text Readability Measures In 1948, Flesch [9] developed a formula that is is considered as one of the oldest and most accurate text readability formulas.

$$R_{Flesch} = 206.835 - 84.6 \cdot \frac{n_{syllables}}{n_{words}} - 1.015 \cdot \frac{n_{words}}{n_{sentences}}$$

The equivalent for Spanish language was developed a few years later by Huerta [7].

$$R_{Huerta} = 206.84 - 60 \cdot \frac{n_{syllables}}{n_{words}} - 102 \cdot \frac{n_{sentences}}{n_{words}}$$

5 Experiments and Results

5.1 Bot Identification

For bot identification subtask we conducted four experiments with five different classifiers (Gradient Boosting, Random Forest, SVM, Logistic Regression, K Nearest Neighbours). The differences between the experiments are more focused on testing the improvement with training data increase, as well as feature set generalization using raw fingerprint n-grams versus statistical diversity measures.

Experiment 1 In Experiment 1 we used character n-grams of user fingerprint described in 4.1. The length of n-grams is a combination of 2, 3 and 4. We can see that some classifiers have fairly similar results (Table 1, column E1). The best classifier is Random Forest for both languages. In this experiment we used the training subsets for English and Spanish separately.

Experiment 2 In Experiment 2 we used the diversity measures calculated on character n-grams of user fingerprint described in 4.1. The length of n-grams is a combination of 1, 2 and 3. The best classifier is Random Forest for both languages. In this experiment we used the training subsets for English and Spanish separately.

Experiment 3 In Experiment 3 (Table 2, column E3) we used the same features as in Experiment 1. The best classifier is Gradient Boosting ensemble for both languages. In this experiment we used the training subsets for English and Spanish combined. Because the features are language independent, we combined training dataset into one, and tested it on both languages. The final model is same for both subsets.

[2] https://spacy.io/

[3] https://unicode.org/Public/emoji/12.0/emoji-test.txt

		E1			E2		
Dataset	Classifier	Precision	Recall	F1	Precision	Recall	F1
English	**GB**	0.9197	0.9153	0.9151	0.9263	0.9234	0.9233
	SVM	0.9174	0.9161	0.9161	0.9253	0.9242	0.9241
	LR	0.8840	0.8750	0.8743	0.9261	0.9242	0.9241
	KNN	$-^{*}$	$-^{*}$	$-^{*}$	0.9284	0.9258	0.9257
	RF	0.9284	0.9218	**0.9215**	0.9293	0.9266	**0.9265**
Spanish	**GB**	0.8666	0.8663	0.8663	0.8429	0.8391	0.8387
	SVM	0.8602	0.8598	0.8597	0.8164	0.8163	0.8163
	LR	0.8663	0.8663	0.8663	0.8510	0.8478	0.8475
	KNN	$-^{*}$	$-^{*}$	$-^{*}$	0.8617	0.8587	0.8584
	RF	0.9115	0.9033	**0.9028**	0.8503	0.8489	**0.8488**

Table 1. Bot classification. Results tested on development dataset. Per language training dataset. $*$ not available due to memory restrictions.

		E3			E4		
Dataset	Classifier	Precision	Recall	F1	Precision	Recall	F1
English	**GB**†	0.9252	0.9242	**0.9241**	0.9330	0.9306	**0.9305**
	SVM	0.9094	0.9081	0.9080	0.9199	0.9177	0.9176
	LR	0.9121	0.9113	0.9112	0.9214	0.9202	0.9201
	KNN	$-^{*}$	$-^{*}$	$-^{*}$	0.9256	0.9242	0.9241
	RF	0.9189	0.9153	0.9151	0.9256	0.9242	0.9241
Spanish	**GB**†	0.8896	0.8880	**0.8879**	0.8512	0.8424	0.8414
	SVM	0.8588	0.8587	0.8587	0.8490	0.8435	0.8429
	LR	0.8478	0.8478	0.8478	0.8473	0.8446	0.8443
	KNN	$-^{*}$	$-^{*}$	$-^{*}$	0.8586	0.8543	**0.8539**
	RF	0.8764	0.8696	0.8690	0.8498	0.8435	0.8428

Table 2. Bot classification. Results tested on development dataset. Combined training dataset. $\dagger$ used as final classifier (E4 for official ranking). $*$ not available due to memory restrictions.

Experiment 4 In Experiment 4 (Table 2, column E4) we used the same features as in Experiment 2. The best classifier for English is Gradient Boosting ensemble and K Nearest Neighbours for Spanish. As in Experiment 3, we combined training dataset into one, and tested it on both languages.

Although a better performance was obtained on separately trained models for two languages (Random Forest, Table 1) with raw features, we opted for Gradient Boosting ensemble which was trained on combined dataset (Spanish portion slightly dropped in performance). The classifier from Experiment 4 was used for the official ranking.

5.2 Gender Identification

For the gender identification subtask we used the same set of classifiers as for bot detection. The results in Table 3 show that Gradient Boosting classifier performed the best

Dataset	Classifier	Precision	Recall	F1
English	**GB**[†]	0.8167	0.8129	**0.8123**
	SVM	0.7782	0.7774	0.7773
	LR	0.7630	0.7629	0.7629
	KNN	0.6054	0.6048	0.6043
	RF	0.7926	0.7919	0.7918
Spanish	**GB**[‡]	0.7062	0.7000	**0.6977**
	SVM	0.6592	0.6587	0.6584
	LR	0.6418	0.6413	0.6410
	KNN	0.5851	0.5848	0.5845
	RF	0.6568	0.6543	0.6530

Table 3. Gender classification. Results tested on development dataset. †, ‡ used as final classifiers.

Dataset	Bot	Gender
English	0.9216	0.7928
Spanish	0.8956	0.7494
Average	0.9086	0.7711

Table 4. Final results on test dataset. Averaged per language.

for both languages. This task was language dependent, so each language had its own model.

5.3 Results on Test Data

The official results are shown in Table 4. Bot detection for English performed with similar results as in our experiments with development set, while for Spanish performed better. Similar improvement was obtained with Spanish dataset for gender identification. The models for the final evaluation are trained on both, training and development sets.

6 Conclusion

We show that automated accounts have less diverse behaviour than genuine user accounts and these measures can help in detecting automated behaviour without diving into language-specific analyses. For the gender identification task we used a standard set of features usually used in stylometry analysis, with the addition of emoji features on a more granular level.

References

1. Ahmed, F., Abulaish, M.: A generic statistical approach for spam detection in online social networks. Computer Communications 36(10-11), 1120–1129 (2013)

2. Bessi, A., Ferrara, E.: Social bots distort the 2016 us presidential election online discussion. First Monday 21(11) (2016)
3. Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., Tesconi, M.: DNA-inspired online behavioral modeling and its application to spambot detection. IEEE Intelligent Systems 31(5), 58–64 (2016)
4. Dadvar, M., Jong, F.d., Ordelman, R., Trieschnigg, D.: Improved cyberbullying detection using gender information. In: Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012). University of Ghent (2012)
5. 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)
6. 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)
7. Fernández Huerta, J.: Medidas sencillas de lecturabilidad. Consigna 214, 29–32 (1959)
8. Ferrara, E., Varol, O., Menczer, F., Flammini, A.: Detection of promoted social media campaigns. In: tenth international AAAI conference on web and social media (2016)
9. Flesch, R., Gould, A.J.: The art of readable writing, vol. 8. Harper New York (1949)
10. Gilani, Z., Wang, L., Crowcroft, J., Almeida, M., Farahbakhsh, R.: Stweeler: A framework for Twitter bot analysis. In: Proceedings of the 25th International Conference Companion on World Wide Web. pp. 37–38. International World Wide Web Conferences Steering Committee (2016)
11. Goswami, S., Sarkar, S., Rustagi, M.: Stylometric analysis of bloggers' age and gender. In: Third international AAAI conference on weblogs and social media (2009)
12. Guess, A., Nagler, J., Tucker, J.: Less than you think: Prevalence and predictors of fake news dissemination on Facebook. Science advances 5(1), eaau4586 (2019)
13. 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)
14. Howard, P.N., Woolley, S., Calo, R.: Algorithms, bots, and political communication in the US 2016 election: The challenge of automated political communication for election law and administration. Journal of Information Technology & Politics 15(2), 81–93 (2018), https://doi.org/10.1080/19331681.2018.1448735
15. Maaten, L.v.d., Hinton, G.: Visualizing data using t-SNE. Journal of machine learning research 9(Nov), 2579–2605 (2008)
16. Messias, J., Schmidt, L., Oliveira, R., Benevenuto, F.: You followed my bot! Transforming robots into influential users in Twitter. First Monday 18(7) (2013)
17. Peersman, C., Daelemans, W., Van Vaerenbergh, L.: Predicting age and gender in online social networks. In: Proceedings of the 3rd international workshop on Search and mining user-generated contents. pp. 37–44. ACM (2011)
18. 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)
19. 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)

20. Sarawgi, R., Gajulapalli, K., Choi, Y.: Gender attribution: tracing stylometric evidence beyond topic and genre. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning. pp. 78–86. Association for Computational Linguistics (2011)
21. Shu, K., Wang, S., Liu, H.: Understanding user profiles on social media for fake news detection. In: 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). pp. 430–435. IEEE (2018)
22. Sichel, H.S.: On a distribution law for word frequencies. Journal of the American Statistical Association 70(351a), 542–547 (1975), https://doi.org/10.1080/01621459.1975.10482469
23. 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)
24. 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)
25. Thelwall, M., Wilkinson, D., Uppal, S.: Data mining emotion in social network communication: Gender differences in MySpace. Journal of the American Society for Information Science and Technology 61(1), 190–199 (2010)
26. Tweedie, F.J., Baayen, R.H.: How variable may a constant be? measures of lexical richness in perspective. Computers and the Humanities 32(5), 323–352 (1998)
27. 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)
28. 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)