Exploring Writing Style Consistency to Timely Identify Heterogeneous Social Bots

Sonia Laudanna[1][0009−0005−1256−7076], Matteo Cardaioli[2], Andrea Di Sorbo[1][0000−0002−3192−739X], Corrado A. Visaggio[1][0000−0002−0558−4450], and Mauro Conti[3][0000−0002−3612−1934]

[1] University of Sannio, Benevento, Italy
{slaudanna, disorbo, visaggio}@unisannio.it
[2] GFT Italy, Milan, Italy
matteo.cardaioli@gft.com
[3] University of Padua, Padua, Italy
mauro.conti@unipd.it

Abstract. Social bots are AI-based algorithms aimed at imitating (and often influencing) the behavior of users on social media. In recent years, bots have been largely used for malicious purposes, like spreading disinformation and conditioning electoral campaigns. To preserve the security and privacy of legitimate users, social media platforms need highly effective solutions to identify bot-driven accounts. Numerous approaches have been proposed to address this problem. However, to better understand how to (i) automatically recognize the specific category a bot belongs to, (ii) early detect bot-driven accounts, and (iii) design platform-independent solutions, further research is needed. In this paper, we consider a stylistic-consistency-based approach for social bot detection and assess whether such an approach can be used to bridge these research gaps. Our results demonstrate that the stylistic-consistency-based approach can (i) identify the specific bot category with an F-measure higher than 95% and (ii) enable near-early detection of bot-driven accounts, achieving high F-measure values when considering a low number of tweets/posts. Though the method is platform-independent, it needs to be trained with platform-specific data to catch the stylistic footprints of bots operating on the particular platform.

1 Introduction

Automated agents on Online Social Networks (OSNs), known as social bots, mimic human behavior for both benign and malicious purposes [12]. Indeed, while some bots are harmless [1], others engage in activities such as promoting businesses [14], harvesting private data [3], spreading misinformation [22] and malware [27]. These bots vary from simple automated accounts to sophisticated AI-enhanced agents [24] and proliferate across OSN platforms. In our previous work [4], we introduced a bot detection method based on analyzing the consistency of writing styles in posts over time. This method assumes that bots maintain a more consistent writing style compared to humans. While effective

on Twitter, its applicability to other platforms is unclear. In addition, effective solutions should (i) recognize specific bot categories for targeted action, (ii) be platform-independent for wider applicability [18], and (iii) detect and neutralize malicious bot accounts early [25].

This paper (i) investigates the approach's ability to model different types of bot writing styles for classification purposes, (ii) determines the necessary amount of data for reliable classification, and (iii) evaluates the approach performance on Reddit, extending beyond its proven success on Twitter. The study finds that the approach achieves high accuracy in identifying the specific bot categories. The approach proves effective for near-early detection of bot-driven accounts on both Twitter and Reddit, exhibiting high F-measure values with a low number of tweets/posts per user and generalizing better than previous works in the literature. In summary, as original contributions of this paper, we (i) explore the effectiveness of writing style consistency-based approaches in automatically recognizing heterogeneous bot types, (ii) empirically study the amount of data required by the stylistic consistency-based models to achieve high performance in social bot identification tasks, and (iii) extend bot prediction from one platform to multiple platforms, experimenting with bot identification on both Twitter and Reddit.

2 Related Work

Stieglitz *et al.* [24] pioneered the classification of social media bot accounts based on their intents and ability to mimic human behavior. Similarly, Cresci *et al.* [7] developed supervised machine learning classifiers to detect fake followers. To overcome the limitations of supervised classifiers and follow the evolution of bots, Chang *et al.* [28] conducted an empirical study on evasion tactics used by Twitter spammers. Others measured various user behavior metrics and compared bot and human values [20], [27]. To address data access limitations, techniques requiring smaller user activity samples and fewer labeled examples of bots and humans have been developed. Examples include classification methods proposed by Chu *et al.* [5], NLP-based detection techniques by Clark *et al.* [6] and BotOrNot [9], and an improved CGAN by Wu *et al.* [26] to enhance detection accuracy.

Ashraf *et al.* [2] proposed a stylometry-based approach combining character-based and emotion-based features, achieving high accuracy in bot classification for the English language. However, social bot detection models are often tailored for specific online social networks, primarily Twitter, with typical features. Pham *et al.* [19] introduced Bot2Vec, leveraging network representation learning for cross-platform bot detection. Cross-platform efforts like BotBuster by Ng and Carley [17] show promise but may require more data for consistent classification. No prior work investigated the effectiveness of an approach relying on the analysis of the stylistic consistency of users in bot classification tasks beyond detection. Our work also demonstrates the feasibility of using such an approach on different platforms while enabling early identification of bot-driven accounts.

3 Study Design

Previous work [18] highlighted the need for further investigation into (i) automatic recognition of specific bot categories, (ii) early-phase bot detection, and (iii) development of platform-independent models. We design our study with the *goal* of investigating whether the approach introduced in [4] can be used to bridge these research gaps. Specifically, we pose three research questions:

- RQ_1: ***To what extent are machine learning models based on stylistic feature consistency able to distinguish different bot types?***
- RQ_2: ***How does the number of considered posts affect the detection effectiveness of models based on stylistic indicator consistency?***
- RQ_3: ***Can stylistic indicator consistency-based models trained on a specific social networking platform be used for detecting bot-driven accounts on other platforms?***

3.1 Context

Table 1. Experimental dataset.

Group name	Description	Accounts	Tweets/ comments	Platform
genuine accounts	human-operated accounts from	3,211	7,896,356	Twitter
fake followers	fake follower accounts from	3,099	195,757	Twitter
traditional spambots	spammer bot accounts from	998	145,085	Twitter
social spambots #1	retweeters of an Italian political candidate from	989	1,610,171	Twitter
social spambots #2	spammers of paid apps for mobile devices from	3,420	427,890	Twitter
social spambots #3	spammers of products on sale at Amazon.com from	462	1,418,619	Twitter
Twitter Total		**12,179**	**11,693,878**	
bot	know Reddit bots extracted from r/autowikibot	343	259,860	Reddit
troll	troll accounts from Reddit's 2017 Transparency Report	153	6,567	Reddit
Reddit Total		**496**	**266,427**	

We used two distinct datasets. The first dataset originates from Cresci *et al.* [8]. It comprises a diverse collection of accounts, including spambots, social spambots, fake followers, and genuine users, sourced from Twitter. This dataset encompasses over 12,179 accounts, both genuine and bot-driven and more than 11.5 million tweets. Given the representation of various bot categories in this dataset, such as spambots and fake followers, we utilized it to address RQ_1-RQ_3. The second dataset is sourced from a study [23] conducted on the Reddit platform. It consists of 343 annotated bot accounts and 156 troll accounts. We selected this dataset for two primary reasons: first, it features bot-driven accounts operating on Reddit, offering distinct characteristics compared to Twitter (e.g., longer posts, subject-based aggregation, voting systems). Second, the inclusion of troll-operated accounts adds complexity to the evaluation, as distinguishing

between the writing styles of bot-driven and troll-operated accounts can be challenging in noisy contexts. However, the Reddit dataset only identifies whether an account is a bot or a troll, without specifying the specific bot categories. Therefore, we used this dataset to address RQ_2-RQ_3 (not RQ_1).

3.2 Metrics Suite

To measure the stylometric properties of text, we used metrics from our previous work [4]. In particular, structural traits were analyzed by treating text as character sequences and counting different types of characters. Lexical traits were measured using vocabulary richness and sentence length. Readability was evaluated using indexes such as Flesch Kincaid Reading Ease, Flesch-Kincaid grade level, Dale-Chall index, Coleman Liau Index, SMOG index, and Linsear Write index. The full list of considered metrics can be found in our replication package[4]. To capture the central tendency and the width of fluctuations in the different users' style traits, the approach [4] models a social network user through the mean and standard deviation values of the stylistic indicators computed on the posts shared by the user.

3.3 Analysis Method

To answer our research questions, we train five different machine learning (ML) classifiers that have been successfully employed in previous work concerning author profiling tasks [21], namely Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), Linear Support Vector Machine (SVM linear), and Support Vector Machine with RBF kernel (SVM RBF). The details of the selected hyperparameters are included in the replication package. Depending on the research question, we used specific configurations for the training, testing and validation phase of the models. Specifically:

- **RQ_1 (distinguish different bot types).** To train our multi-class classifiers, we perform a stratified nested 10-fold cross-validation to split the Twitter dataset in training and test set, and a 10-fold inner cross-validation for hyper-parameters selection and model validation. Moreover, to assess the generalizability of the approach, we apply a *Leave-One-Botnet-Out* (LOBO) validation [11]. With this strategy, we selectively exclude a category of bots from the training set, using it only in the test set. The LOBO validation is specifically conducted to estimate the detection effectiveness and resilience of the models in a real scenario where the detection approach is likely to deal with new types of bots (i.e., not represented in the training dataset).
- **RQ_2 (post number influence).** The detection effectiveness of models based on stylistic indicators consistency is tested using a gradually increasing number of posts for each user. Concretely, we extract the metrics explained in Section 3.2 and compute the mean and standard deviation of such metrics by considering different windows containing a fixed number of posts

[4] https://github.com/papersubs/socialBotDetection

randomly chosen for each user of the Twitter and Reddit datasets. In particular, we consider windows of 1, 10, 20, 50, 100, and all the posts (i.e., *full*). To verify the influence of the number of posts considered on the prediction performance, we perform a nested-cross-fold-validation on our ML models, with a 10-fold inner cross-validation and a 10-fold outer cross-validation.

– **RQ$_3$ (cross-dataset assessment).** To investigate the transferability of the stylistic feature consistency-based approach, we train the models on one dataset (i.e., Twitter) and test them on the other (i.e., Reddit). Further, given Twitter's limitations on the number of characters that can be used per post, we evaluate both a normalized version of the metrics by the number of characters (i.e., *cross normalized*) and a non-normalized one. In both cases, we apply 10-fold cross-validation on the training dataset (i.e., Twitter) for hyperparameter selection.

The performance of machine learning models is evaluated using the F1 score with a micro average.

4 Results

4.1 *RQ$_1$* (distinguish different bot types)

Table 2 presents the F1 score (micro) results for both all bot classes tested and the LOBO tests. The outcomes reveal that the multiclass classification using the entire Twitter dataset achieves an F1 score (micro) ranging from 86% with the DT classifier to 96% with the SVM RBF classifier. In line with prior work [4], the SVM RBF is the model achieving the best performance not only in discerning between human-operated and bot-driven accounts but also in identifying the specific category of bots behind the bot-driven ones.

Further insights into ML model diagnostic abilities are provided in the replication package. Among the classifiers, DT and RF models demonstrate promising results in detecting genuine accounts but struggle with traditional spambot classification (i.e., a Precision of 62.7% for the DT model and a Recall of 64.7% for the RF model). LR and SVM linear models excel in identifying genuine accounts but show lower recall for fake followers. Notably, the SVM RBF classifier achieves superior performance across all categories, with Precision and Recall exceeding 93% for all classes, particularly excelling in identifying social spambots. Looking deeper into the types of errors made by the best-performing classifier (i.e., SVM RBF), it is possible to observe that only about 3% of actual genuine accounts are wrongly identified as bot-driven and less than 3% of actual bot-driven accounts are classified as human-operated.

Fake followers are the types of bots that are more frequently misclassified as genuine. Indeed, about 4% of fake followers are wrongly classified as legitimate, while, for the other types of bot accounts, only 1.5% is erroneously recognized as human-operated. Finally, almost 4% of the traditional spambots are detected as belonging to the fake follower category. These results confirm the high detection effectiveness of the SVM RBF model that is able, in the vast majority of the

Table 2. F1 score (micro) of ML models, in nested 10-fold cross-validation, to distinguish different bot types (RQ1).

		DT	RF	LR	SVM Linear	SVM RBF
Experiment	**# Posts**	**F1 (Micro)**	**F1 (Micro)**	**F1 (Micro)**	**F1 (Micro)**	**F1 (Micro)**
Twitter dataset	Full	0.86 (0.01)	0.89 (0.01)	0.94 (0.00)	0.94 (0.00)	0.96 (0.00)
LOBO (fake follower)	Full	0.62	0.47	0.64	0.64	0.36
LOBO (social spambot)	Full	0.89	0.88	0.87	0.87	0.81
LOBO (traditional spambot)	Full	0.81	0.65	0.79	0.81	0.87

cases, to predict if an account is a bot- or human-driven one and also to recognize with high accuracy the type of bot driving an automated account. Considering the LOBO validation, in the case of the social spambot class used only in the testing set, the F1 score (micro) for each ML model is above 80% (see Table 2). In the case of the traditional spambot class used only in the testing set, the value of the F1 score (micro) measure ranges from 65% to 87%. An important remark to note is that the performance of the classification models substantially drops in terms of F1 score (micro) when the fake follower class is not represented in the training set. This could be explained by the fact that spambot tweets (or traditional spambot tweets) usually have more complex content than tweets from fake followers, affecting the classification outcome.

To better understand the differences in stylistic traits between the fake followers and the other bot types, we tested the following null hypothesis:

H_0*: the distributions of the metrics, calculated for each type of bot, are equal.*

through the Mann-Whitney test (fixing p-value to 0.05) [16] for all the metrics discussed in Section 3.2. In addition, to quantitatively assess the extent to which these groups are different, we used Cliff's delta [13]. In the replication package, we report the results of this investigation that has been carried out for the various bot categories: (i) fake followers, (ii) spambots, and (iii) traditional spambots. Fake followers exhibit a substantially different writing style compared to the other types of bots. In particular, in the comparison between fake followers and traditional spambots, it is possible to note that fake followers typically generate posts that are easier to read (based on the Flesh Kincaid Grade Level), contain more digits, punctuation characters, URLs, and unique words, with lower variability (i.e., standard deviation) than traditional spambots. Conversely, compared to more advanced spambots, fake followers tend to utilize more words, words with multiple syllables, prepositions, blank spaces, and lowercase words, while employing fewer special characters.

4.2 RQ_2 (post number influence)

Table 3 displays the F1 score (micro) achieved by various ML models for each dataset alongside the number of posts included. For the Twitter dataset, only 50 posts suffice to exceed an F1 score (micro) of 90% across all models. The SVM RBF model emerges as the most effective for early detection, requiring ten tweets per user to achieve an F1 score ≥ 0.90, whereas other models necessitate at least double the posts per user to reach a similar F1 score. Tree-based ML

Table 3. F1 score (micro) of ML models, in nested 10-fold cross-validation related to post number influence (RQ2).

		DT	RF	LR	SVM Linear	SVM RBF
Social	**# Posts**	**F1 (Micro)**	**F1 (Micro)**	**F1 (Micro)**	**F1 (Micro)**	**F1 (Micro)**
Twitter	1	0.54 (0.01)	0.55 (0.01)	0.55 (0.01)	0.55 (0.01)	0.67 (0.01)
Twitter	10	0.68 (0.02)	0.74 (0.01)	0.86 (0.01)	0.84 (0.01)	0.90 (0.01)
Twitter	20	0.77 (0.01)	0.82 (0.01)	0.92 (0.01)	0.91 (0.01)	0.95 (0.01)
Twitter	50	0.90 (0.01)	0.90 (0.01)	0.95 (0.01)	0.95 (0.00)	0.97 (0.01)
Twitter	100	0.93 (0.01)	0.94 (0.00)	0.95 (0.01)	0.95 (0.01)	0.97 (0.00)
Twitter	Full	0.86 (0.01)	0.89 (0.01)	0.94 (0.00)	0.94 (0.00)	0.96 (0.00)
Reddit	1	0.71 (0.05)	0.79 (0.08)	0.78 (0.06)	0.79 (0.07)	0.79 (0.06)
Reddit	10	0.90 (0.04)	0.92 (0.05)	0.92 (0.02)	0.92 (0.02)	0.90 (0.05)
Reddit	20	0.94 (0.03)	0.94 (0.02)	0.92 (0.03)	0.90 (0.05)	0.91 (0.06)
Reddit	50	0.95 (0.03)	0.95 (0.03)	0.95 (0.04)	0.93 (0.04)	0.95 (0.04)
Reddit	100	0.94 (0.04)	0.96 (0.04)	0.96 (0.04)	0.93 (0.02)	0.95 (0.04)
Reddit	Full	0.81 (0.03)	0.86 (0.05)	0.86 (0.05)	0.87 (0.05)	0.88 (0.04)

models (decision tree and random forest) are the slowest to converge, requiring at least 50 tweets per user to obtain an F1 score ≥ 0.90. When experimenting with the Reddit dataset, the number of posts sufficient to achieve an F1 score ≥ 0.90 drops to ten for all models. This may be due to the characteristics of the Reddit posts (e.g., they may be longer, as no character limitations are imposed) that can favor the ML models to more easily isolate the stylistic footprints characterizing the different types of users by considering a lower number of posts. We employed the RF algorithm to rank the importance of variables (stylistic consistency metrics) in the classification problem. This analysis allowed us to identify the most influential features for predicting bot types on Twitter and distinguishing between bot and troll accounts on Reddit. Notably, features like NumberOfSpecialCharactersAvg, AutomatedReadabilityIndexAvg, and NumberOfNumbersAvg emerge as highly important and relevant across both Twitter and Reddit datasets, demonstrating their generalizability for identifying bots across diverse social networking platforms. Complete MDA results for all features are available in the replication package.

4.3 RQ_3 (cross-dataset assessment)

To answer RQ3, we select the Twitter dataset as the traning dataset and the Reddit dataset as the testing one. We perform experiments on two variations of posts engineering: with normalization (i.e., the content of each post is divided by the value of the *NumberOfTotalCharacters* feature) and without normalization (i.e., using the features extracted from posts as they are). This choice is motivated by the fact that, on the Twitter platform, the maximum number of characters per tweet is set to 280, while for Reddit posts, no character limitations are imposed. Table 4 shows that ML algorithms achieve F1 scores (micro) ranging from 42% to 58% when making predictions on Reddit data with post normalization, but do not exceed 47% in the other case. Further analysis using the Mann-Whitney test with Cliff's delta value reveals key differences between bot types on Reddit and Twitter platforms (see replication package). For exam-

Table 4. F1 score (micro) of ML models, in nested 10-fold cross-validation related to cross-dataset assessment (RQ3).

		DT	RF	LR	SVM Linear	SVM RBF
Experiment	**# Posts**	**F1 (Micro)**	**F1 (Micro)**	**F1 (Micro)**	**F1 (Micro)**	**F1 (Micro)**
Cross (normalized)	Full	0.53	0.43	0.54	0.58	0.42
Cross	Full	0.44	0.42	0.45	0.47	0.4

ple, the posts produced by bots on the Reddit platform have higher values (with *large* effect size) relating to the readability metrics (*AutomatedReadabilityIndexAvg* and *FleschKincaidGradeLevelAvg*) and the usage of special characters (i.e., *NumberOfSpecialCharactersAvg*) while the tweets produced by bots on the Twitter platform have higher values (with *large* effect size) for the following features: *LengthOfWordsStd (d=0.5018), NumberOfUppercaseWordsStd (d=0.5594) and, VocabularyRichnessStd (d=0.4801)*. The experimented ML models can effectively detect bot-driven accounts on both Twitter and Reddit platforms, when trained with platform-specific data. However, cross-platform application results in low classification performance due to differences in writing style richness and variability among bots on different platforms. This underscores the need for platform-specific training examples to achieve high classification performance.

5 Threats to Validity

Threats to construct validity relate to possible imprecision in the measurements we performed. In our study, we estimate different factors that could be insufficient to comprehensively shape the writing style of a social media user. To mitigate this, we selected metrics previously used to address similar challenges [15, 21].

Threats to internal validity concern confounding factors that could affect our results. Inaccuracies in identifying human and bot-driven users in our datasets could be a significant confounding factor. We addressed this by using publicly available datasets with manually verified accounts [8, 23].

Threats to conclusion validity concern the relationship between treatment and outcome. To draw our conclusions, we adopted appropriate non-parametric statistical procedures. We used the Mann-Whitney test and Cliff's delta effect size measure to investigate significant differences in stylistic consistency indicators among bot types and platforms.

Threats to external validity concern the generalizability of the findings. Our datasets primarily include specific bot types and platforms, potentially limiting generalization. To address this, future research should replicate our study at a larger scale, considering more diverse bots and platforms.

6 Conclusions

Social bots are AI-based algorithms aimed at imitating the behavior of users on social media. Effective solutions to identify bot-driven accounts are crucial,

yet missing, particularly ones adaptable across platforms and capable of early detection of malicious bots. We demonstrated that a machine learning-based approach analyzing users' writing style consistency [4] effectively distinguishes between the different types of genuine and bot-driven accounts. It achieves high classification performance with minimal data and is flexible enough to perform well on diverse social networking platforms such as Twitter and Reddit.

In the future, we aim to expand the analysis to additional bot types and social media platforms. We also plan to enhance the approach by considering new metrics capturing traits of advanced bots and incorporating NLP techniques to recognize language patterns [10] used by different bot types.

References

1. Arin, E., Kutlu, M.: Deep learning based social bot detection on twitter. IEEE Transactions on Information Forensics and Security **18**, 1763–1772 (2023)
2. Ashraf, S., Javed, O., Adeel, M., Iqbal, H., Nawab, R.M.A.: Bots and gender prediction using language independent stylometry-based approach. CLEF (Working Notes) **100** (2019)
3. Bilge, L., Strufe, T., Balzarotti, D., Kirda, E.: All your contacts are belong to us: automated identity theft attacks on social networks. In: Proceedings of the 18th international conference on World wide web. pp. 551–560 (2009)
4. Cardaioli, M., Conti, M., Di Sorbo, A., Fabrizio, E., Laudanna, S., Visaggio, C.A.: It's a matter of style: Detecting social bots through writing style consistency. In: 2021 International Conference on Computer Communications and Networks (ICCCN). pp. 1–9. IEEE (2021)
5. 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)
6. Clark, E.M., Williams, J.R., Jones, C.A., Galbraith, R.A., Danforth, C.M., Dodds, P.S.: Sifting robotic from organic text: a natural language approach for detecting automation on twitter. Journal of computational science **16**, 1–7 (2016)
7. Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., Tesconi, M.: Fame for sale: Efficient detection of fake twitter followers. Decision Support Systems **80**, 56–71 (2015)
8. Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., Tesconi, M.: The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In: Proceedings of the 26th international conference on world wide web companion. pp. 963–972 (2017)
9. 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 (2016)
10. Di Sorbo, A., Panichella, S., Visaggio, C.A., Di Penta, M., Canfora, G., Gall, H.C.: Exploiting natural language structures in software informal documentation. IEEE Transactions on Software Engineering **47**(8), 1587–1604 (2019)
11. Echeverrï£¡ a, J., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Stringhini, G., Zhou, S.: Lobo: Evaluation of generalization deficiencies in twitter bot classifiers. In: Proceedings of the 34th annual computer security applications conference. pp. 137–146 (2018)

12. Ferrara, E., Varol, O., Davis, C., Menczer, F., Flammini, A.: The rise of social bots. Communications of the ACM **59**(7), 96–104 (2016)
13. Hess, M.R., Kromrey, J.D.: Robust confidence intervals for effect sizes: A comparative study of cohen'sd and cliff's delta under non-normality and heterogeneous variances. In: annual meeting of the American Educational Research Association. vol. 1. Citeseer (2004)
14. Li, H., Mukherjee, A., Liu, B., Kornfield, R., Emery, S.: Detecting campaign promoters on twitter using markov random fields. In: 2014 IEEE International Conference on Data Mining. pp. 290–299. IEEE (2014)
15. López-Anguita, R., Montejo-Ráez, A., Díaz-Galiano, M.C.: Complexity measures and pos n-grams for author identification in several languages. Tech. rep., Retrieved 7/2/2020, from shorturl. at/kFIOQ (2018)
16. McKnight, P.E., Najab, J.: Mann-whitney u test. The Corsini encyclopedia of psychology pp. 1–1 (2010)
17. Ng, L.H.X., Carley, K.M.: Botbuster: Multi-platform bot detection using a mixture of experts. arXiv preprint arXiv:2207.13658 (2022)
18. Orabi, M., Mouheb, D., Al Aghbari, Z., Kamel, I.: Detection of bots in social media: a systematic review. Information Processing & Management **57**(4), 102250 (2020)
19. Pham, P., Nguyen, L.T., Vo, B., Yun, U.: Bot2vec: A general approach of intra-community oriented representation learning for bot detection in different types of social networks. Information Systems p. 101771 (2021)
20. Pozzana, I., Ferrara, E.: Measuring bot and human behavioral dynamics. Frontiers in Physics **8**, 125 (2020)
21. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the author profiling task at pan 2013. In: CLEF Conference on Multilingual and Multimodal Information Access Evaluation. pp. 352–365. CELCT (2013)
22. Shao, C., Ciampaglia, G.L., Varol, O., Yang, K.C., Flammini, A., Menczer, F.: The spread of low-credibility content by social bots. Nature communications **9**(1), 1–9 (2018)
23. Skowronski, J.: Identifying trolls and bots on reddit with machine learning (part 2) (2019), https://towardsdatascience.com/identifying-trolls-and-bots-on-reddit-with-machine-learning-709da5970af1
24. Stieglitz, S., Brachten, F., Ross, B., Jung, A.K.: Do social bots dream of electric sheep? a categorisation of social media bot accounts. arXiv preprint arXiv:1710.04044 (2017)
25. Valliyammai, C., Devakunchari, R.: Distributed and scalable sybil identification based on nearest neighbour approximation using big data analysis techniques. Cluster Computing **22**, 14461–14476 (2019)
26. Wu, B., Liu, L., Yang, Y., Zheng, K., Wang, X.: Using improved conditional generative adversarial networks to detect social bots on twitter. IEEE Access **8**, 36664–36680 (2020)
27. Yan, G., Chen, G., Eidenbenz, S., Li, N.: Malware propagation in online social networks: nature, dynamics, and defense implications. In: Proceedings of the 6th acm symposium on information, computer and communications security. pp. 196–206 (2011)
28. Yang, C., Harkreader, R., Gu, G.: Empirical evaluation and new design for fighting evolving twitter spammers. IEEE Transactions on Information Forensics and Security **8**(8), 1280–1293 (2013)