LDK 2023

WIDISBOT: Widget to analyse disinformation and content spread by bots

Jose Manuel Camacho *
Institute of
Mathematical Sciences
(ICMAT-CSIC)

Luis Perez-Miguel
Institute for
Physical and Information
Technologies
(ITEFI-CSIC)

David Arroyo
Institute for
Physical and Information
Technologies
(ITEFI-CSIC)

Abstract

The increasing prevalence of bots poses a significant challenge for maintaining the integrity of online information. Bot campaigns have been deployed for both economic scams and political interference, making it necessary to develop a system to detect these agents and analyze their behavior. We present a scalable application designed to identify bots and to buttress the investigation of disinformation campaigns. Our intention is to provide professionals without technical expertise with an effective tool to identify and analyze content generated by bots. This will enable researchers from diverse backgrounds to study bot activity, fostering an interdisciplinary understanding of the strategies these agents use to spread disinformation, and the characteristics of their discourse. We illustrate how to use the application through a case study on COVID-19.

1 Introduction

In a world characterized by an increasing globalization and the rapid dissemination of information, many decisions are influenced by publicly accessible information obtained through online sources. In 2021, more than 50% of Twitter's users were obtaining news directly from the platform (Pew Research Center, 2021). Individuals who rely on social media for news tend to exhibit reduced engagement with news and possess limited knowledge regarding a wide range of current events (Pew Research Center, 2020). This creates an exploitable opportunity for malicious actors to manipulate public opinion or deceive unsuspecting users through disinformation, posing a threat to the 16th Sustainable Development Goal of the United Nations, which aims for an inclusive and peaceful society (Bontcheva et al., 2020).

One of these malicious agents are bots, software programs that can mimic human behavior on social networks like Twitter. They have played a significant role in the dissemination of low credibility content (Shao et al., 2018), and their presence continues to grow within the discourse of democratic processes (Pastor-Galindo et al., 2020). Moreover, they can be combined with Large Language Models to generate counterfeit news and fabricate speech that resembles that of a human (De Angelis et al., 2023). Given the limited effectiveness of current methods for detecting non-human content (Pegoraro et al., 2023), it is crucial to adopt a different perspective. Instead of solely focusing on the accuracy of the content, an alternative approach is to identify bots based on their behavior, which can be inferred from the analysis of their metadata. Based on bot detection techniques, it is also possible to expose disinformation campaigns that have the potential to influence critical decision-making processes.

WIDISBOT [1] has been developed to address the challenge of scrutinizing the dissemination of disinformation by bots in Twitter. This tool employs a scalable machine learning model and enables the analysis of bot discourse in tweets, making comparisons with human users participating in the same public conversations. This discourse analysis comprises the examination of sentiment, hashtags, and the usage of the most shared URLs or hashtags. Built using Streamlit [2], the primary goal of this widget is to offer professionals with non-technical expertise an effective means for examining how bots propagate disinformation. It empowers them to contribute to research on these agents and enhance the field with insights from diverse disciplines. By enhancing interdisciplinary research, we facilitate the development of information consumption security frameworks and contribute to safeguard digital societies.

*Corresponding author: josemanuel.camacho@icmat.es

[1] The application is available at: https://github.com/jmcamachor1/WIDISBOT

[2] https://streamlit.io/

2 Related works

Research on bot detection has significantly increased over the last decade, leading to the development of various methods, with supervised learning being the most widely adopted approach (Cresci, 2020). A conspicuous example of a supervised method is demonstrated in (Yang et al., 2020), where the account's metadata is utilized to construct a scalable detector. Another popular alternative for bot detection is unsupervised learning, which does not rely on labeled datasets. An illustrative instance of this method is given in (Mazza et al., 2019), where the identification of bot accounts is constructed upon the analysis of the temporal patterns of retweeting behavior. One popular method for modeling bot behavior involves generating a string, similar to a DNA chain, that can encode different aspects of bot behavior (Cresci et al., 2017). This modeling can be exploited from both supervised and unsupervised learning methods. An additional alternative is to employ an adversarial approach (Najari et al., 2022), which mitigates the impact of evasion techniques on bot detection.

Bot detection models have been integrated into user-friendly software, making them accessible to individuals with no technical expertise. One notable example is Botometer (Sayyadiharikandeh et al., 2020), which enables users to predict the likelihood of an account being a bot by leveraging over 1200 features. Otherwise, Bot Detective (Kouvela et al., 2020) offers a web service powered by an explainable method for detecting bots. BotSlayer introduces a system with a dashboard to visualize the users who are sharing content that matches a given Twitter query (Hui et al., 2019, 2020). The system provides various metrics and allows content filtering based on entities such as hashtags, user handles, and links. One of these metrics focuses on assessing the likelihood of an account being a bot, which can be accomplished using different rules or bot detection models. Combining BotSlayer with Hoaxy enables the analysis of the spread of disinformation associated with bots and their corresponding fact-checking responses (Shao et al., 2016).

Our approach, WIDISBOT, facilitates the comparison of discourse between bots and humans within a specific conversation on Twitter. Users can input either a Twitter query or tweets IDs, enabling further analysis of tweets datasets. WIDISBOT offers an interface to visualize disparities in discourse between automated and genuine users by applying sentiment and words frequency analysis. Additionally, WIDISBOT supports in-depth examination of fabricated content that is propagated by these entities.

3 Application description

This section presents an overview of the application's functionalities and the machine learning (ML) models empowering them. Initially, we outline the application capabilities, followed by a description of the models. When analyzing tweets through the various functionalities, the input format requires Tweet Objects obtained via the Twitter API, and the related User Object representing the tweet author.

3.1 Functionalities

Below, we describe the application functionalities:

- *Data extraction (DE).* It enables the retrieval of tweets by connecting to the API. Therefore, valid credentials are necessary. These can be for any version of the Twitter API (v1.1, v2). The retrieved data is then normalized in the structure of v1.1 Tweet Objects and User Objects. In particular, the user may extract tweets by ID, or via search containing a certain keyword, hashtag or URL on a specific date. This functionality is limited by Twitter API restrictions and rate limits. The generated dataset can then be used as an entry to any other WIDISBOT functionality.

- *Monitoring (M).* It identifies which of the input tweets were generated by bots or humans. Additionally, it plots the probability distribution that the accounts that posted those tweets were bots, as well as the proportion of those accounts that were labelled as bots or humans and the number of tweets produced by each account type.

- *Forensics (F).* Given the accounts' usernames, it computes the likelihood of them being bots, allowing the results to be presented in an aggregated manner.

- *Sentiment analysis (SA).* It computes the sentiment of the input tweets, displaying the human/bot sentiment distribution in both a discrete (positive-negative-neutral) and continuous fashion.

	Test datasets					
	botwiki-verified	*cresci-rtbust- 2019*	*gilani-2017*	*kaiser*	*cresci-stock-2018*	*midterm-2018*
Light (v1.1)	**.990**	.613	.631	**.944**	.631	**.964**
Light (v.2)	.975	.518	.580	.936	.653	.947
Botometer v3	.922	**.625**	**.689**	.829	**.756**	.958

Table 1: AUC scores of the bot detection models on different datasets [3]. The *botwiki-verified* is formed through merging datasets *botwiki-2019* and *verified-2019*.

- *Hashtag analysis (HA).* It allows the visualization of the most frequently used hashtags by both humans and bots within the input tweets. This functionality is not case-sensitive, as bots may utilize variations of the same hashtag to promote diverse content.
- *Wordcloud (W).* It provides a visualization of the 25 most frequent words on the tweets shared by bots and humans.
- *Analysis of spread sources (ASS).* It displays the most shared URLs by bots and humans. It is connected to the Wayback Machine [4] to retrieve the content from deleted websites, as content spread by bots is often removed after a certain time. The app also provides access to Media Bias Fact Check [5] to determine the bias of a media and if it is a non-reliable source.
- *Analysis of discourse around hashtags (ADH).* It enables the utilization of the functionalities *M*, *SA*, *HA*, *ASS* on tweets that contain a specific hashtag, allowing for the analysis of how the discourse surrounding the given hashtag is influenced by both bots and humans.

The classification of input accounts as bots or humans is conducted using a threshold specified by the user. A higher threshold leads to a more cautious approach by the model in determining which accounts are classified as bots. It is advisable to utilize a threshold of at least 0.51, although higher thresholds can be employed for a more conservative analysis. Additionally, the application enables users to download files with the results of the various functionalities for further analysis, either manually or in another application.

3.2 Machine Learning models

We provide details about the bot detection and sentiment analysis models integrated into the widget, powering the previous functionalities.

[4] https://archive.org/web/
[5] https://mediabiasfactcheck.com/

Bot detection The widget utilizes the *Light* model from (Antenore et al., 2022) if the input Tweet objects are in Twitter API v1.1 format. However, if the input tweets are in API v2 format, we employ an adapted version of the model that does not consider features inaccessible in API v2 but available in v1.1. These models offer scalability, requiring only a Tweet object to forecast whether an account is a bot. Table 1 demonstrates their effectiveness in detecting various types of bots. Furthermore, they achieve comparable performance to Botometer v3 (Yang et al., 2019), a widely used method for Twitter bot detection (Rauchfleisch and Kaiser, 2020). Additionally, since the model solely relies on language-agnostic features, it can predict tweets irrespective of their language.

Sentiment analyzer The app employs VADER (Hutto and Gilbert, 2014) as the sentiment analysis model. VADER utilizes a lexicon to assign scores to each word, which are subsequently combined using five rules that consider grammatical and syntactical aspects. The output is a unidimensional continuous metric (y) ranging from -1 (most negative) to 1 (most positive). To categorize y discretely, we use the thresholds provided by the authors: positive if $y > 0.05$, negative if $y < -0.05$, and neutral if $-0.05 \leq y \leq 0.05$. VADER is computationally efficient and scalable. Additionally, it performs well across various domains, particularly in analyzing microblogging content. In fact, according to (Ribeiro et al., 2016), it is an effective method for predicting three-class sentiment in social network messages.

4 Case study

This section displays how the application could be used to study bots' role on a potential disinformation campaign. For illustrative purposes, we have selected a set of 527 tweets used in experiments in (Antenore et al., 2022) from 7th February 2020 that contain the words 'Trump' and 'deathtoll', and their subvariants. These tweets were produced at the start of the COVID-19 pandemic when there

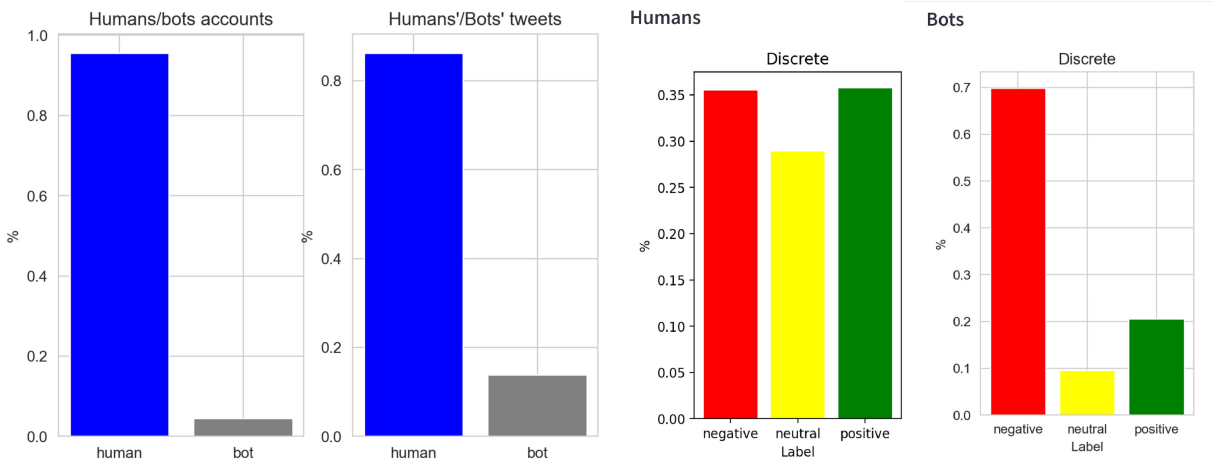

Figure 1: Screenshots of WIDISBOT output. (Left) Proportion of accounts in the subset labelled as bots/humans and the fraction of tweets produced by each type. (Right) Sentiment distribution in human/bots tweets.

was still much uncertainty about the health crisis. We aim to display how to use the widget to study whether bots intended to promote certain content by taking advantage of the crisis situation. We follow the steps below to carry out the tweets' analysis:

1. *Analysis of bot presence.* Utilising the *M* functionality, we examined the proportion of tweets produced by bots compared to humans. In Figure 1 (left) we observe that a smaller number of bots produced a larger proportion of the total tweets than humans, an indication that bots are interested to promote content in this conversation.

2. *Checking differences in sentiment.* Another indication of bot activity may be differences in the sentiment distribution between bots and humans. We used the *SA* functionality to determine if any differences were present. Specifically, as depicted in Figure 1 (right), we observed substantial discrepancies, evidence about the different content that bots and humans are sharing.

3. *Checking differences between hashtags.* Through the *HA* functionality, we examined how hashtags were used by both groups of accounts. The results for the 10 most used hashtags by bots and humans are depicted in Figure 2. We observed that bots used more hashtags and, while there was a stair-like shape in the human case, the bots had several hashtags with the same number of occurrences. This may be an indication that bots are promoting their content using multiple hashtags in the same tweets.

4. *Studying tweets with a certain hashtag.* We studied hashtag $\#deathtoll$ as it was highly shared by bots, but not at the same rate as the first six hashtags, and it was not among the most frequently used hashtags by humans. We utilized the *ASH* functionality and discovered that only one human and one bot posted tweets with the hashtag. However, the bot produced 44 tweets while the human produced only one. Furthermore, we examined the URLs shared by the bot on these tweets, observing that it shared 34 times the same URL.

5. *Analysis of the most shared URLs.* We browsed the most shared URL by the bot, finding out that it is no longer available. To check the content, we used the *ASS* functionality and retrieved the website content during the period when the tweet was produced. Figure 3 displays the website. It can be observed that some content is advertised, such as how to survive without medication or publicity about masks. Hence, we have uncovered that the identified bot was disseminating content that could potentially contribute to disinformation during the COVID-19 pandemic.

5 Discussion

This paper introduces WIDISBOT, a widget specifically developed to identify automated accounts on

[5]Datasets are accesible in https://botometer.osome.iu.edu/bot-repository/datasets.html

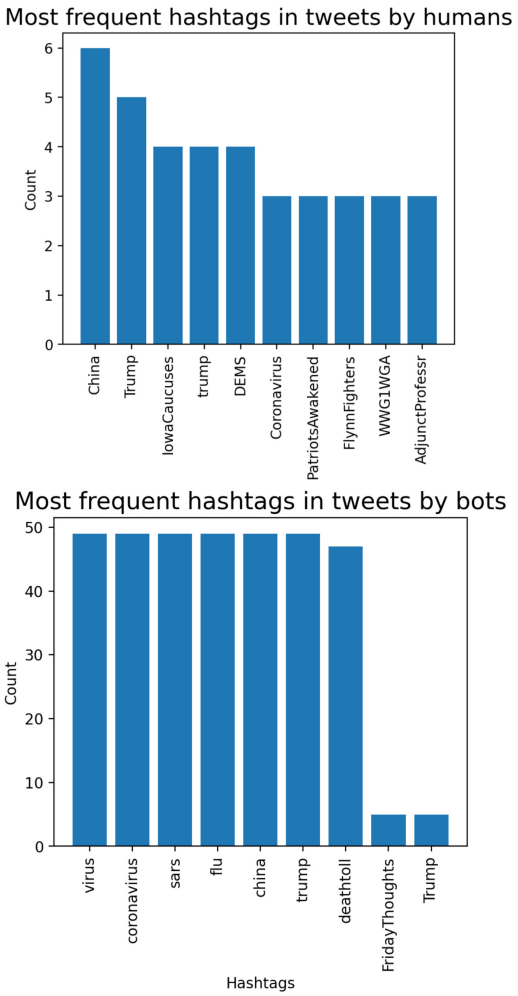

Figure 2: Ten most shared hashtags by bots and humans.

Twitter and analyze the content they promote in comparison to human users. By offering various functionalities, our aim is to provide application users with a comprehensive perspective on the information disseminated by both genuine users and bots. Additionally, we present a use case demonstrating how the widget can be utilized to uncover campaigns that potentially propagate disinformation during COVID-19 pandemic.

We have developed a user-friendly system utilizing Streamlit, which features an intuitive interface specifically designed for non-technical users, such as journalists and social scientists engaged in researching the spread of disinformation by bots. The widget demonstrates scalability and serves as an effective tool for examining disparities in content between human and automated accounts, and it is compatible with different Twitter API access. Future extensions of the widget will consist of incorporating more ML models to analyze other aspects of bot discourse, such as determining whether certain content constitutes any form of hate speech. Furthermore, it will be integrated with other applications that concentrate on identifying specific forms of misinformation, such as (Arroyo Guardeño et al., 2023), in order to bolster the versatility of WIDISBOT within specific contexts.

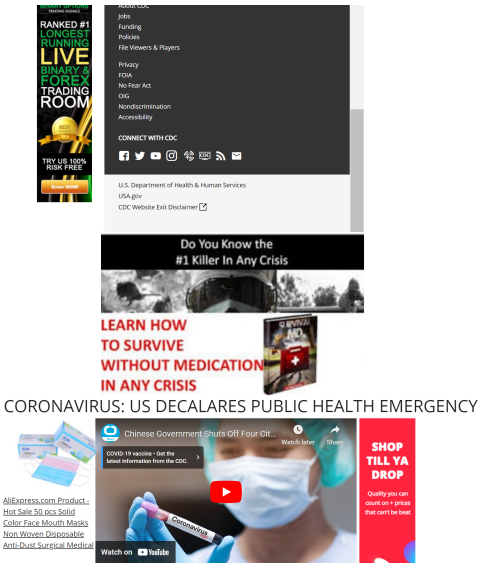

Figure 3: Screenshot of the most shared website by bots in tweets with hashtag $\#deathtoll$, accessed through the Wayback Machine.

Acknowledgements

This work was supported by the Spanish Ministry of Science program PID2021-124662OB-I00; a fellowship from "la Caixa" Foundation (ID 100010434), whose code is LCF/BQ/DI21/11860063; and Grant PLEC2021-007681 (project XAI-DisInfodemics) funded by MCIN/AEI/ 10.13039/501100011033 and by European Union NextGeneration EU/PRTR; and BBVA Foundation project AMALFI.

References

Marzia Antenore, Jose Manuel Camacho Rodriguez, and Emanuele Panizzi. 2022. A Comparative Study of Bot Detection Techniques with an Application in Twitter COVID-19 Discourse. *Social Science Computer Review*, page 08944393211073733.

David Arroyo Guardeño, Gómez Espés Alberto Degli Esposti, Sara, Santiago Palmero Muñoz, and Luis Pérez-Miguel. 2023. On the design of a misinformation widget (Ms. W) against cloaked science. In *NSS 2023: 17th International Conference on Network and System Security*.

Kalina Bontcheva, Julie Posetti, Denis Teyssou, Trisha Meyer, Sam Gregory, Clara Hanot, and Diana Maynard. 2020. Balancing act: Countering digital disinformation while respecting freedom of expression. *Geneva, Switzerland: United Nations Educational, Scientific and Cultural Organization*.

Stefano Cresci. 2020. A decade of social bot detection. *Communications of the ACM*, 63(10):72–83.

Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2017. Social fingerprinting: detection of spambot groups through DNA-inspired behavioral modeling. *IEEE Transactions on Dependable and Secure Computing*, 15(4):561–576.

Luigi De Angelis, Francesco Baglivo, Guglielmo Arzilli, Gaetano Pierpaolo Privitera, Paolo Ferragina, Alberto Eugenio Tozzi, and Caterina Rizzo. 2023. ChatGPT and the Rise of Large Language Models: The New AI-Driven Infodemic Threat in Public Health. *Available at SSRN 4352931*.

Pik-Mai Hui, Kai-Cheng Yang, Christopher Torres-Lugo, and Filippo Menczer. 2020. Botslayer: Diy real-time influence campaign detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 980–982.

Pik-Mai Hui, Kai-Cheng Yang, Christopher Torres-Lugo, Zachary Monroe, Marc McCarty, Benjamin D Serrette, Valentin Pentchev, and Filippo Menczer. 2019. Botslayer: real-time detection of bot amplification on twitter. *Journal of Open Source Software*, 4(42):1706.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Maria Kouvela, Ilias Dimitriadis, and Athena Vakali. 2020. Bot-detective: An explainable twitter bot detection service with crowdsourcing functionalities. In *Proceedings of the 12th International Conference on Management of Digital EcoSystems*, pages 55–63.

Michele Mazza, Stefano Cresci, Marco Avvenuti, Walter Quattrociocchi, and Maurizio Tesconi. 2019. Rtbust: Exploiting temporal patterns for botnet detection on twitter. In *Proceedings of the 10th ACM conference on web science*, pages 183–192.

Shaghayegh Najari, Mostafa Salehi, and Reza Farahbakhsh. 2022. Ganbot: a gan-based framework for social bot detection. *Social Network Analysis and Mining*, 12:1–11.

Javier Pastor-Galindo, Mattia Zago, Pantaleone Nespoli, Sergio López Bernal, Alberto Huertas Celdrán, Manuel Gil Pérez, José A Ruipérez-Valiente, Gregorio Martínez Pérez, and Félix Gómez Mármol. 2020. Spotting political social bots in Twitter: A use case of the 2019 Spanish general election. *IEEE Transactions on Network and Service Management*, 17(4):2156–2170.

Alessandro Pegoraro, Kavita Kumari, Hossein Fereidooni, and Ahmad-Reza Sadeghi. 2023. To ChatGPT, or not to ChatGPT: That is the question! *arXiv preprint arXiv:2304.01487*.

Pew Research Center. 2020. Americans who mainly get their news on social media are less engaged, less knowdgeable.

Pew Research Center. 2021. News consumption across social media in 2021.

Adrian Rauchfleisch and Jonas Kaiser. 2020. The false positive problem of automatic bot detection in social science research. *PloS one*, 15(10):e0241045.

Filipe N Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. 2016. Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5:1–29.

Mohsen Sayyadiharikandeh, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2020. Detection of novel social bots by ensembles of specialized classifiers. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 2725–2732.

Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2016. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th international conference companion on world wide web*, pages 745–750.

Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2018. The spread of low-credibility content by social bots. *Nature communications*, 9(1):1–9.

Kai-Cheng Yang, Onur Varol, Clayton A Davis, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2019. Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, 1(1):48–61.

Kai-Cheng Yang, Onur Varol, Pik-Mai Hui, and Filippo Menczer. 2020. Scalable and generalizable social bot detection through data selection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 1096–1103.