Social Network Analysis and Mining (2025) 15:7
https://doi.org/10.1007/s13278-025-01410-5

ORIGINAL ARTICLE

Dissecting a social bot powered by generative AI: anatomy, new trends and challenges

Salvador Lopez-Joya[1,2] · Jose A. Diaz-Garcia[1,2] · M. Dolores Ruiz[1,2] · Maria J. Martin-Bautista[1,2]

Received: 23 September 2024 / Revised: 14 November 2024 / Accepted: 13 December 2024
© The Author(s) 2025

Abstract
The rise of social networks has transformed communication, information sharing and entertainment, but it has also facilitated the rise of harmful activities such as the spread of misinformation, often through the use of social bots. These automated accounts that mimic human behaviour have been implicated in significant events, including political interference and market manipulation. In this paper, we provide a comprehensive review of recent advances in social bot detection, with a particular focus on the role of generative AI and large language models. We present a new categorisation scheme for bots that aims to reduce class overlap while maintaining generality. In addition, we analyse the most commonly used datasets and state-of-the-art classification techniques, and through user profile-based measures, we use Explainable Artificial Intelligence (XAI) and data mining techniques to uncover factors that contribute to bot misclassification. Our findings contribute to the development of more robust detection methods, which are essential for mitigating the impact of malicious bots on online platforms.

Keywords Bot detection · Generative AI · Social networks analysis · XAI · Data mining

1 Introduction

The proliferation of social media has undeniably transformed our daily lives, becoming an integral part of our communication with family and friends, a source of information on various topics, a platform for work, and a means of entertainment. The success of social media is closely tied to the rise of Artificial Intelligence (AI), as the vast data generated from these platforms has become an invaluable resource for companies and researchers. This data is reshaping human-AI interactions and revolutionizing the ways in which humans interact both with one another and with algorithms, leading to remarkable advancements (Shin 2023).

However, this success has also fueled harmful activities, such as the deliberate spread of misinformation, which can have serious consequences for various sectors of society that can be more suitable to be affected by orchestrated campaigns (Shin 2024). Many nations have raised concerns about foreign interference in their electoral processes and social movements, often orchestrated by other countries or organizations (Linvill and Warren 2020; Nisbet et al. 2021; Kennedy et al. 2022; Freelon et al. 2022). A significant portion of this disinformation is propagated by social bots, automated accounts that mimic human behaviour on social networks, creating and sharing content while interacting with unsuspecting users who are typically unaware that they are engaging with artificial entities. This issue presents a challenge where artificial intelligence can play an important role in mitigation.

Social bots have a significant impact on critical societal functions. For instance, during the US 2016 elections, bots played a central role and possibly contributed to the victory of Donald Trump. This influence extends beyond politics; in 2017, it was estimated that 15% of Twitter accounts were bots (Varol et al. 2017), while in 2019, 11% of Facebook

✉ Salvador Lopez-Joya
slopezjoya@ugr.es

Jose A. Diaz-Garcia
jagarcia@decsai.ugr.es

M. Dolores Ruiz
mdruiz@decsai.ugr.es

Maria J. Martin-Bautista
mbautis@decsai.ugr.es

[1] Department of Computer Science and A.I, University of Granada, C. Periodista Daniel Saucedo Aranda, 18014 Granada, Spain

[2] Research Centre for Information and Communications Technologies, C. Periodista Rafael Gómez Montero, 18014 Granada, Spain

Published online: 10 March 2025

accounts were bots (Zago et al. 2019). Moreover, an alarming 71% of users discussing US stocks on Twitter in 2017 were identified as bots (Cresci et al. 2019a; Cresci 2020). These examples highlight the pervasive presence and potential influence of bots in social media, reinforcing the urgent need for advanced bot detection methodologies to safeguard the integrity of online discourse and information dissemination as we can see in Fig. 1.

Motivated by these reasons, extensive research is being conducted in the area of social bot detection. Our aim is to provide a compendium of the latest advances to help researchers in their further investigations. In line with Ferrara (2023), our research also highlights generative AI techniques and discusses the challenges, trends and new avenues this technology opens for bot detection. Furthermore, this research presents a detailed data-driven analysis of the features and behavioural patterns useful for detecting and categorising bots, proposing unique insights not covered by previous research. Our work contributes to the current state of the art in several ways:

- We provide a thorough analysis of existing literature and theoretical foundations in bot detection, focusing on integrating large language models (LLMs) and generative AI.
- The proposal of a new categorization schema that covers most of the bots observed in social networks and focused on reducing the overlap between classes while maintaining abstraction.
- We provide a thorough review and analysis of the most commonly used datasets and classification approaches for bot detection, helping researchers to select the most appropriate resources for their work.
- We have conducted an analysis using XAI to identify the features that influence bot misclassification. Through this analysis, we aim to uncover the underlying factors that contribute to bot misclassification, thereby improving the effectiveness of detection methods.

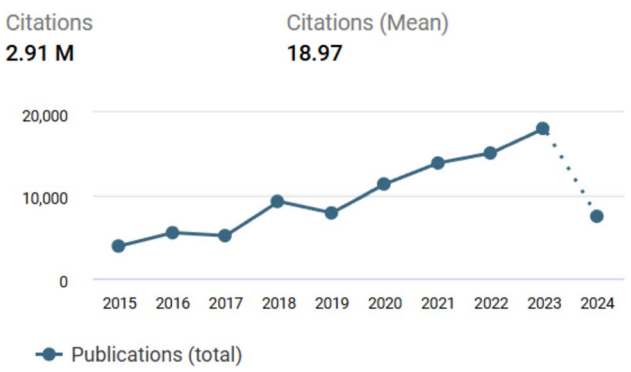


Fig. 1 Social bot detection publications. Source dimensions.ai (https://www.dimensions.ai/)

The paper is organized as follows: next section describes the methodology used to conduct our review. Section 3 offers a comprehensive and detailed review of current trends in bot detection, with a special subsection dedicated to the impact of generative AI on bot detection and the most widely used datasets. Section 4 presents a detailed analysis using explainable artificial intelligence techniques to uncover why bot classification can sometimes be challenging. In Sect. 5, we present a new classification schema for bot categorization based on their inherent characteristics. Section 6 provides a point-by-point review of our findings aligned with our research questions. Finally, Sect. 7 presents the conclusions of our research.

2 Methodology

In this paper we have relied on PRISMA methodology (Moher et al. 2010), enabling us to conduct a robust and exhaustive review tailored to our research needs. This methodology is based on four main stages: identifying research questions, defining eligibility criteria, establishing a search strategy, and selecting relevant studies.

2.1 Research questions

As we live in a world that is very changeable and almost each day the artificial intelligence area is evolving with the incursion of generative AI, the aim of our research is to provide a robust starting point for further innovations, putting together the state-of-the-art, a comprehensive method to categorize bots, the analysis of the influence of new technologies in bot detection and creation, and finally, providing a point-by-point analysis of why classification systems fail in bot categorization. To do this, our paper tries to solve the following research questions:

- **RQ1:** What are the current state-of-the-art in terms of bot detection?
- **RQ2:** Which features contribute to the complexity of bot detection in social media?
- **RQ3:** How does the emergence of generative artificial intelligence impact both the detection and creation of social bots?
- **RQ4:** Can all different categories of bots be grouped into one single categorization schema?

2.2 Eligibility criteria

Our eligibility criteria were designed to select studies that provided substantial insights into bot detection methodologies, particularly focusing on those utilizing advanced techniques. Specifically, we included studies that:

- Addressed social bot detection within the context of microblogging platforms, with a specific emphasis on Twitter.
- Were published from 2019 onwards to ensure relevance to recent advancements in AI and bot detection.
- Utilised methodologies grounded in machine learning, natural language processing, or similar data science disciplines.
- Provided insights into the challenges and trends in bot detection, including the influence of new technologies.
- Studies available in English.

We also include references to articles that have been considered necessary for a proper understanding of the key points in the detection of social bots and the techniques they use, even if they do not meet any of these criteria.

2.3 Search strategy

To ensure that our review reflected the latest research, we set a cut-off date of 2019, a year marked by significant advances in bot detection technologies, most notably the introduction of BERT (Bidirectional Encoder Representations from Transformer (Devlin et al. 2018)).

To identify relevant studies, we used a range of search terms and queries, including "*bot detection in social media*", "*(Twitter OR 𝕏 OR microblogging) bot detection*", "*identifying bots in (Twitter OR 𝕏 OR microblogging)*", "*social bot detection*", "*large language models social bot detection*", etc, along with keywords associated with key data science methodologies like "*machine learning*" and "*natural language processing*".

These searches were performed across various databases including Scopus, Web of Science, and Google Scholar. Other databases, such as IEEE Xplore, SpringerLink, and ACM, are implicitly included in our review since their journals and publications are already indexed in databases like Google Scholar and Scopus. These primary databases cover the vast majority of published research, allowing us to conduct a robust and comprehensive review. Additionally, we manually reviewed reference lists of relevant articles to identify additional studies that met our inclusion criteria.

2.4 Study selection

Following the initial search using our predefined criteria and keywords, we screened the titles and abstracts of the identified articles. Articles that appeared relevant based on their title and abstract were selected for full-text review. During this stage, we applied our inclusion criteria to ensure the selected studies aligned with our research objectives. Furthermore, we cross-referenced the bibliographies of selected articles to identify additional relevant studies that might have been missed during the initial search.

Throughout this process, we have consulted and verified all potential inclusions. This collaborative approach ensured thorough consideration of each study against the predefined inclusion criteria.

A total of 453 articles were retrieved from the initial search of the databases. After removing duplicates and reviewing their titles and abstracts to ensure they met our inclusion criteria, 70 articles remained: 36 articles containing a new bot detection approach or dataset, and 34 articles related to social bot detection, including 9 articles identified through manual reference searches. The entire process carried out in the review is shown in the PRISMA flow chart in Fig. 2.

3 Bot detection in social media

While there is a general consensus among authors that a bot on social networks involves some degree of automation, there is no widely accepted definition that comprehensively covers all relevant aspects of these accounts. This ambiguity is due to the rapid evolution of technology and the uncertainties in attributing specific characteristics to bots. One aspect under debate is the level of automation required for an account to be classified as a bot, given the existence of accounts that are partially automated also called cyborgs. In addition, the definition of bot have been based on its similarity to human behaviour, which varies widely among researchers, with some emphasis to the extent that a bot mimic human actions in social contexts. Furthermore, the study of bots extends across multiple disciplines, with computer scientists focusing more on technical attributes and social scientists emphasising the broader social implications (Cresci 2020).

We can give a more or less restrictive definition depending on how we value these characteristics. For example, Abokhodair et al. (2015) defines a social media bot as any program that behaves like a human in a social space. In contrast, Morstatter et al. (2016) defines it as any account within social media that is controlled by software. Others, such as Yang et al. (2020), Assenmacher et al. (2020), note that an account can only be partially automated. One of the most restrictive definitions is provided by Ferrara et al. (2016), which describes a social media bot as a program that interacts with humans in a social environment, automatically produces content, and aims to mimic and potentially alter human behaviour.

In this study we follow the definition given in Lopez-Joya et al. (2023) which states the following: *"a social media bot is an account that is automated enough to produce content*

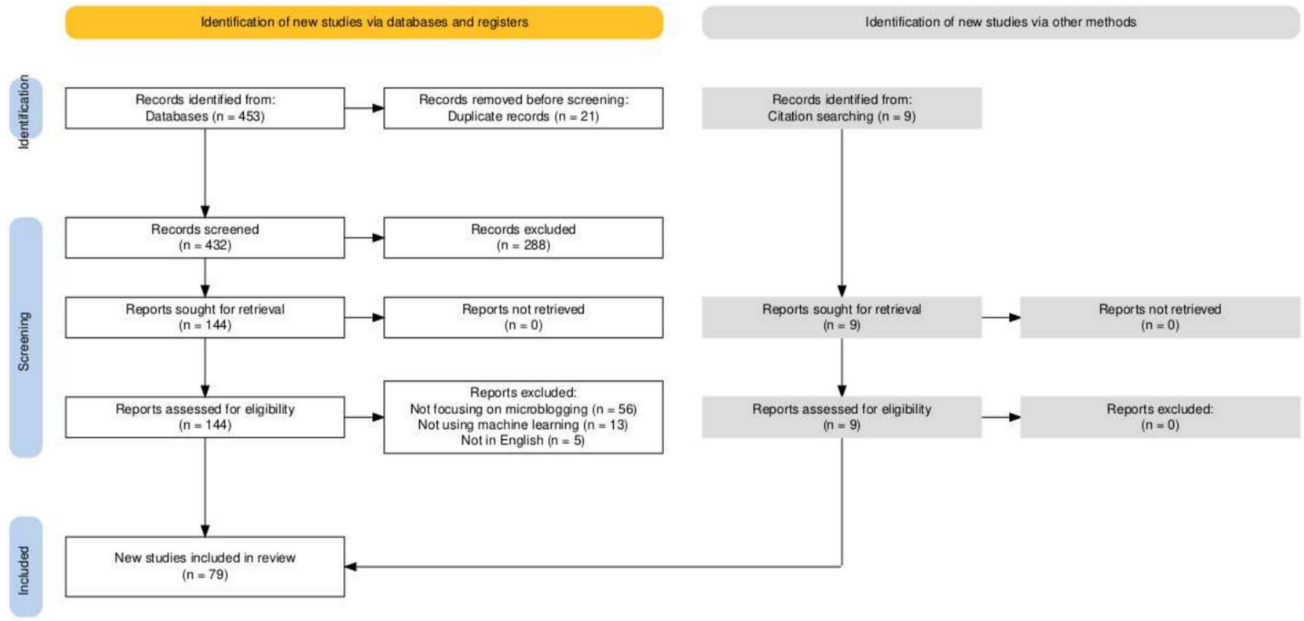


Fig. 2 Prisma flow diagram

and/or interact with other accounts within a social media context."

3.1 Current trends in bot detection

The history of bots and bot detectors is a cyclical struggle. When a new way of detecting bots is born, a new generation of more sophisticated bots is also born with the goal of avoiding that new method (Cresci et al. 2019b). Early approaches to the problem aimed to create a general-purpose classifier using supervised learning techniques and focusing on individual accounts. They assumed that legitimate individual accounts could be distinguished from bot accounts based on their (Cresci 2020) characteristics. But, with the advent of more sophisticated bots, the researchers had to rethink that assumption.

Trying to detect simple bots is a relatively overpowered task for the most popular types of bots and there are many detectors in the literature that do this very well. The problem comes when these bots behave in more sophisticated ways. In Cresci et al. (2017) we can see that sophisticated bots can mimic human behaviour so well that even humans have trouble distinguishing a legitimate account from an account that is not. In Cresci et al. (2017) the authors conduct an experiment with volunteers by hand sorting 4428 accounts between the classes: spam bot, genuine, and unable to classify. The results show that human reviewers obtained less than 24% accuracy on this task.

In recent years, this problem is more present than ever. With great language generation models like GPT (Generative Pre-trained Transformer), the development of different image generation tools and deep fakes, bot accounts can look even more like legitimate ones. Making the line between what is real and what is fake blurred more than ever. This is one of the reasons why there is a growing trend in recent years to focus bot detectors on studying groups of bots rather than individual accounts (Cresci 2020).

Another avenue being explored in recent years is the inclusion of BERT (Devlin et al. 2018) in bot detectors. When the Google paper came out, it dominated the state of the art of natural language processing. Some authors have started using BERT as a basis for building text-based bot detectors. Examples of its use can be seen in Dukić et al. (2020), where this language representation model is used to directly extract features from tweet content for further use in a deep neural network, or in Heidari and Jones (2020), where it is used for sentiment analysis of tweets for further classification. For this reason, there is a growing trend to use this language representation model in bot detectors.

Most of the state-of-the-art techniques used by researchers can be grouped into the following categories. Within these categories, we will give some notable examples of bot detectors that focus on or can be applied to social networks from recent years. The compilation and key points of these examples are shown in Tables 1 and 2.

3.1.1 Graph-based

Given our focus on social networks, it is intuitive to use graph-based techniques to detect bots. These methods

Table 1 Bot detection studies following inclusion criteria

Paper	Date	Social network	Category	Datasets	Keys
Yang et al. (2020)	2020-04-03	Twitter	Feature based	caverlee, varol-icwsm, cresci-17, pronbots, celebrity, vendor-purchased, botometer-feedback, political-bots, gilani-17, cresci-rtbust, cresci-stock, mid-term-18, botwiki, verified	Random Forest, SHAP Values, generality
Wu et al. (2023c)	2023-06-14	Twitter	Graph based	TwiBot-20, TwiBot-22	Heterophily-aware contrastive learning, supervised contrastive learning
Najari et al. (2022)	2021-11-14	Twitter	Feature based	Cresci-17 (partially)	Custom GAN with LSTM
Sayyadiharikandeh et al. (2020), Yang et al. (2022)	2022-11-01	Twitter	Feature based	varol-icwsm, cresci-17, pronbots, celebrity, vendor-purchased, botometer-feedback, political-bots, gilani-17, cresci-rtbust, cresci-stock, botwiki, mid-term-2018, astroturf, kaiser	Specialised Random Forests, ensemble
Lingam et al. (2019)	2019-11-01	Twitter	Graph based	Cresci-17, social honeypot	Deep Q-network architecture
Wu et al. (2021)	2021-01-09	Sina Weibo	Feature based	–	Deep neural networks, active learning, new features
Dialektakis et al. (2022)	2022-05-31	Twitter	Feature based	24 datasets mostly from Botometer repository	Conditional adversarial learning
Heidari et al. (2020)	2020-11	Twitter	Feature based	Cresci-17	Word embeddings, multiple neural networks, split into similar profiles
Mazza et al. (2019)	2019-06-26	Twitter	Feature based	Cresci-rtbust	temporal activity of retweets, LSTM autoencoder, hierarchical density-based clustering
Abou Daya et al. (2020)	2020-03	Twitter	Graph based	CTU-13	Anomaly detection, centrality measures, supervised and unsupervised learning
Li et al. (2023a)	2023-07-01	Twitter	Graph based	Own dataset	Community detection, feature engineering

involve modelling users and their interactions as a graph in order to examine certain characteristics. In such graphs, the nodes represent entities within the network (e.g. user accounts), while the edges represent connections between these nodes (e.g. interactions between users). Various features such as centrality measures, community detection techniques, and graph neural networks (GNNs) are used by researchers to identify bots. There have been several notable studies in recent years that highlight the use of these techniques.

In Guo et al. (2021) they propose a combination between the use of BERT and convolutional graph networks, thus realising a model with large-scale pre-training and transductive learning. In Li et al. (2023a) we can highlight the use of community detection techniques and machine learning combined with feature engineering, claiming the state of the art in their field. In Lingam et al. (2019) they address the problem by designing a deep Q-network architecture incorporating Deep Q-Learning. They use each of a user's social attributes as states and define the agent's movement from one state to another as an action.

A unique approach is employed in Rout et al. (2020), which uses URL-based features together with a learning automata-based algorithm by integrating a trust computation model. They compute direct trust and indirect trust based on Bayes' theorem and Dempster-Shafer theory (DST) respectively. A different approach is given in Wu et al. (2023c) where they use a heterophily-aware contrastive learning method that is able to differentiate neighbour representations of heterophilic relations adaptively. In addition, they use supervised contrastive learning to aggregate class-specific information.

In Peng et al. (2024) they propose an unsupervised and interpretable framework for bot detection called UnDBot. Their method consists of three separate modules: first, they model three new relationships: Posting Type Distribution, Posting Influence and Follow-to-follower Ratio to construct a weighted social multi-relational graph, in the second phase they build a two-dimensional encoding tree based on the

Table 2 Bot detection studies following inclusion criteria

Paper	Date	Social network	Category	Datasets	Keys
Guo et al. (2021)	2022-01	Twitter	Graph based	Cresci-rtbust, botometer-feedback, gilani, cresci-stock-2018, midterm	BERT, convolutional graph networks
Grimme et al. (2023)	2023	Twitter	Feature based	3.6 million tweets collected in Pohl et al. (2022)	Siamese Neural Networks, time-series-based interpretation
Ayoobi et al. (2023)	2023-09-05	LinkedIn	Feature based	Own dataset	Fake profiles detection, embeddings, tags and subtags for sections and subsections
Feng et al. (2024)	2024-07-04	Twitter	Graph based	TwiBot-20, TwiBot-22	LLMs with context injection, account metadata, user description and interactions, ensemble
Dukić et al. (2020)	2020-10	Twitter	Feature based	PAN	BERT, deep neural network
Heidari and Jones (2020)	2020-10	Twitter	Feature based	Cresci-17	BERT, sentiment analysis
Arin and Kutlu (2023)	2023	Twitter	Feature based	varol-2017, cresci-2017, botometer-feedback, caverlee-2011	3 LSTMs + fully connected layer
Peng et al. (2024)	2024-04-25	Twitter	Graph based	Cresci-15, cresci-17, pronbots-2019, botwiki-2019	Multi-relational graph construction, user community division, community binary classification
Wu et al. (2023a)	2023-03-25	Twitter	Feature based	Cresci-17	Behavioral patterns, time series
Wu et al. (2023b)	2023-05	Twitter	Feature based	Cresci-17	Embeddings, triplet learning
Yang et al. (2023)	2023-04-30	Twitter	Feature based	Vendor-19, TwiBot-20	Federated adversarial contrastive knowledge distillation, GAN-based, cross-lingual

principle of structural entropy minimisation, and finally they perform a binary community classification using the stationary distribution and entropy of each community.

The last work we highlight in this section is the study conducted in Abou Daya et al. (2020). This research introduces a bot detection system that relies on anomaly detection techniques. It utilizes various centrality measures, including In-Degree (ID) and Out-Degree (OD), In-Degree Weight (IDW) and Out-Degree Weight (ODW), Betweenness Centrality (BC), Local Clustering Coefficient (LCC), and Alpha Centrality (AC). The system employs a two-phase machine learning process that combines both supervised and unsupervised learning to determine whether an account is a bot.

3.1.2 Feature-based

Feature-based methods are the most widespread in the literature. These methods try to use the information that we can find both in the metadata of the account and in the content of the text written by the user. In this category we will include most of the methods based on machine learning and deep learning techniques. These methods are divided into three categories:

- Account-based. They use the user's account information as features or to infer new ones, e.g., account age, username length, number of retweets, number of followers, or follower growth rate.
- Content-based. They use information from the content of tweets as features, for example, the number of URLs, the number of hashtags, the sentiment or the length of the tweet.
- Hybrids. They use a combination of features from the user's account and their content.

Within this group of techniques a possible new trend can be observed. The need to anticipate new adaptations of malicious bots is mentioned in Cresci (2020); Cresci et al. (2021), for which the use of GANs (Generative Adversarial Networks) is proposed. A GAN is a deep learning framework consisting of two neural networks, one generative and one discriminative, which participate in a competitive process to generate realistic data. Genetic algorithms are used to synthetically produce new sophisticated bots in the first network and, in addition, to generate robustness against possible new real sophisticated bots (Cresci et al. 2021). We give below some examples of work done in recent years on feature-based bot detectors.

In Mazza et al. (2019) a very interesting method, belonging to the content-based methods, is proposed. It is based on the temporal activity of retweets between accounts on Twitter. They transform the retweet time series data into latent feature vectors with an LSTM (Long Short-Term Memory)

autoencoder and cluster these vectors using a hierarchical density-based algorithm. Accounts in large clusters with malicious retweeting patterns are identified as bots. In Wu et al. (2021), the authors use a combination of deep neural networks with active learning. They obtain 30 features divided into 4 categories: metadata-based, interaction-based, content-based, and timing-based, and include 9 new features proposed by them.

The study presented in Hayawi et al. (2022) proposes a bot detection algorithm called DeeProBot (Deep Profile-based Bot detection framework), in this study they use a LSTM that has as input variables of different types (numerical, binary and text), they also use the description of the profiles of the users transforming them into embeddings and using them in the input of this network.

The studies Sayyadiharikandeh et al. (2020) and Yang et al. (2022) propose a public Twitter bot detection tool called Botometer, which has become one of the most important in the field. The method behind this tool consists of dividing the features into six categories: user profile, friends, network, temporal, content and language, and sentiment, and by means of several specialised Random Forest classifiers, they use an ensemble to give a final bot score. Currently this tool only works with historical Twitter data due to new API policies.

In Dialektakis et al. (2022) they propose a method to detect bots based on GANs. This study falls into the category of hybrid methods, as they use both features from tweets, such as temporal patterns and sentiment analysis, and also include features from the user's account. The authors use two GAN models to create realistic synthetic bots of multiple types that are added to the dataset. This provides robustness to the model and the ability to proactively detect evolving bots of various types. Finally, they use a Random Forest to give a final classification. Another study using GANs is presented in Najari et al. (2022). In this case, the strategy followed by the authors is to introduce a LSTM between the generator and the discriminator in order to reduce the limitation of convergence that the Sequence Generative Adversarial Net has.

Another study that relies on LSTMs for bot detection is presented in Arin and Kutlu (2023). They use a deep learning architecture with three LSTMs and a merging fully connected (FC) layer at the end, also they explore three learning schemes to train each component effectively.

The approach taken in Fazil et al. (2021) involves the use of an attention-aware deep neural network model. The authors use a Bidirectional Long Short Term Memory (BiLSTM) and a Convolutional Neural Network (CNN), modelling profile, temporal and activity information as sequences for BiLSTM and content for CNN, giving a final classification.

In Yang et al. (2020) the authors pursue a bot detection model that is capable of being effective in most datasets. For this they make a detailed study of the most important datasets and features for bot detection using SHAP Values (SHapley Additive exPlanations). They use some datasets for training and others for testing and use Random Forest combined with the selection of the most relevant features.

Federated learning has also been used to address the problem of bot detection; in the work presented in Yang et al. (2023), they present FedACK, a framework for social bot detection that combines federated adversarial learning, contrastive learning, and knowledge distillation. It enables cross-lingual and cross-model bot detection by using a GAN-based architecture, where a global generator extracts and distils global data distribution knowledge into local models, while local discriminators and generators allow for customised model design and data enrichment.

In Wu et al. (2023a) it is considered the timestamps available in the dataset they use to extract behavioural patterns using time series. The authors consider seasonality and use shapelets representation to maximise the information obtained.

The approach followed in Wu et al. (2023b) is to build a bot detection system that refines content embeddings using triplet learning. Inspired by Sentence BERT, they improve raw embeddings by maximising the distance between bots and real users. Their system, called BOTTRINET, uses a symmetric multilayer perceptron as its embedding network, with a triplet loss function that adjusts parameters to optimise bot detection. They also introduce a triplet selector algorithm for robust sample selection, and use account embeddings to capture long-term behaviour, integrating classifiers such as SVM and Random Forest for final bot detection.

As a final study we have the one made in Heidari et al. (2020). Their work consists of the development of a bot detector based on the use of word embeddings of the text of tweets. The authors use GLOVE (Global Vectors) (Pennington et al. 2014) and ELMO (Embeddings from Language Models) (Sarzynska-Wawer et al. 2021) for a contextualized semantic representation of the tweet text. In the next phase of the detector, eight neural networks are trained based on four features: age, personality, gender and education. Due to the combination of using word embeddings of the tweet text and these features, we can classify this study as a hybrid method. The authors claim that training networks to classify between humans and bots by dividing the dataset according to similar profiles increases the classification accuracy. In the last phase of the detector, a final model is implemented that has as input the values resulting from the different networks of the previous step. The authors experiment with different architectures for the final model to see which one

gives the best results, with the Feedforward Neural Network (FNN) being the winner.

3.1.3 Crowdsourcing

Crowdsourcing techniques are in disuse and rely on human participation when detecting bots in social networks. Some works such as Wang and Hamilton (2012) using online platforms have been proposed for detection, but it was more effective in the early days of social networks. Now it is a time-consuming, expensive and not scalable technique.

To this we must add one of the most problematic types of bots we have mentioned in the classification: cyborgs. Accounts that are sometimes actually legitimate, which makes detection of these bots by humans a really difficult task.

3.2 Bots in the generative artificial intelligence era

To contribute to the latest advances, challenges, and trends in bot detection, we have analyzed how generative AI (Koonchanok et al. 2024), specifically LLMs, are influencing the field. LLMs are revolutionizing the world of NLP and social network analysis (Jain et al. 2024; ShabaniMirzaei et al. 2023), leading us to review their application in the context of bot detection.

LLMs and generative models as GPT have demonstrated an unprecedented ability to generate human-like text. These models are trained on diverse datasets containing vast amounts of internet text, enabling them to produce coherent and contextually relevant content. LLMs can mass-produce text very similar to that written by a person and people have very easy access to this technology. This advancement opens the door to malicious uses, such as the creation of fraudulent computer science assessments and tasks, a problem that some researchers are currently addressing (Richards et al. 2024). This impact has prompted researchers to explore how these models influence society, particularly focusing on the cognitive mechanisms users engage when interacting with AI-generated content and how these mechanisms affect their ability to distinguish credible information from misinformation (Shin et al. 2024). In the context of social bots, Generative Artificial Intelligence can further amplify these risks, as bots can leverage this technology to produce and disseminate misinformation or propaganda on a large scale (Li et al. 2023b).

An example of the use of these large language models can be seen in the LinkedIn network where they are used to generate fake profiles. In Ayoobi et al. (2023) they talk about the growth in the number of fake profiles on this network and how the lack of Lindekin verification encourages the problem. The authors also propose a method for identifying these profiles based on embeddings and using tags for the sections and subsections of the profile. In Grimme et al. (2023) the authors highlight the existence of campaigns generated with LLMs on social networks and talk about how many of the detection techniques developed to date are obsolete. Also they provide an approach based on Siamese Neural Networks integrating the classification results into a time-series-based interpretation of the topic-based campaign detection mechanism. These bot-driven campaigns are unmasked in Yang and Menczer (2023), where the authors reveal a network of 1,140 Twitter accounts leveraging LLMs to generate and disseminate machine-created content. This study illustrates how LLM-enabled malicious bots operate, offering insights into their content generation strategies and the potential evolution toward a greater sophistication. The accounts were identified through accidental tweets that exposed their bot nature, further verified through social network analysis. They also provide a dataset of these bot accounts, creating a valuable resource for further research into the behavior and characteristics of LLM-powered bots.

One of the most interesting indicators of the reach of this technology is that exists a social network where all the content is generated by LLMs. This network is called Chirper[1] and it is a free access social network where you can create your own bot, describe how you want them to behave and their description and they automatically start interacting on this social network as if they were a real person (Fig. 3).

However, the current state of the art of LLMs also creates opportunities to leverage their capabilities in bot detection. The rise of bot detection using large language models LLMs has roots in the early application of language models like BERT for bot detection and classification tasks. In Heidari et al. (2021), for example, the authors utilized BERT for fake news classification, guided by bot activity detection during the COVID-19 pandemic. Although preliminary, the study yielded promising results and provided a starting point for analyzing how bots spread misinformation and how models like BERT can aid in their detection. In a similar approach focusing on sentiment features, Heidari and Jones (2020) proposed a system capable of handling bot-generated content, achieving 94% accuracy on the Cresci dataset. More recently, BERT has remained a powerful tool for more sophisticated bot detection applications, as demonstrated in Harrag et al. (2021), where it was used to detect social media text generated by GPT-2 models. This research achieved high accuracy in generated text detection through a transfer learning approach that also leveraged neural network models such as BI-GRU and BI-LSTM. The effectiveness of pretrained LLMs has also been highlighted in studies such as Sallah et al. (2024), where the authors demonstrate how a simple fine-tuning

[1] https://chirper.ai/.

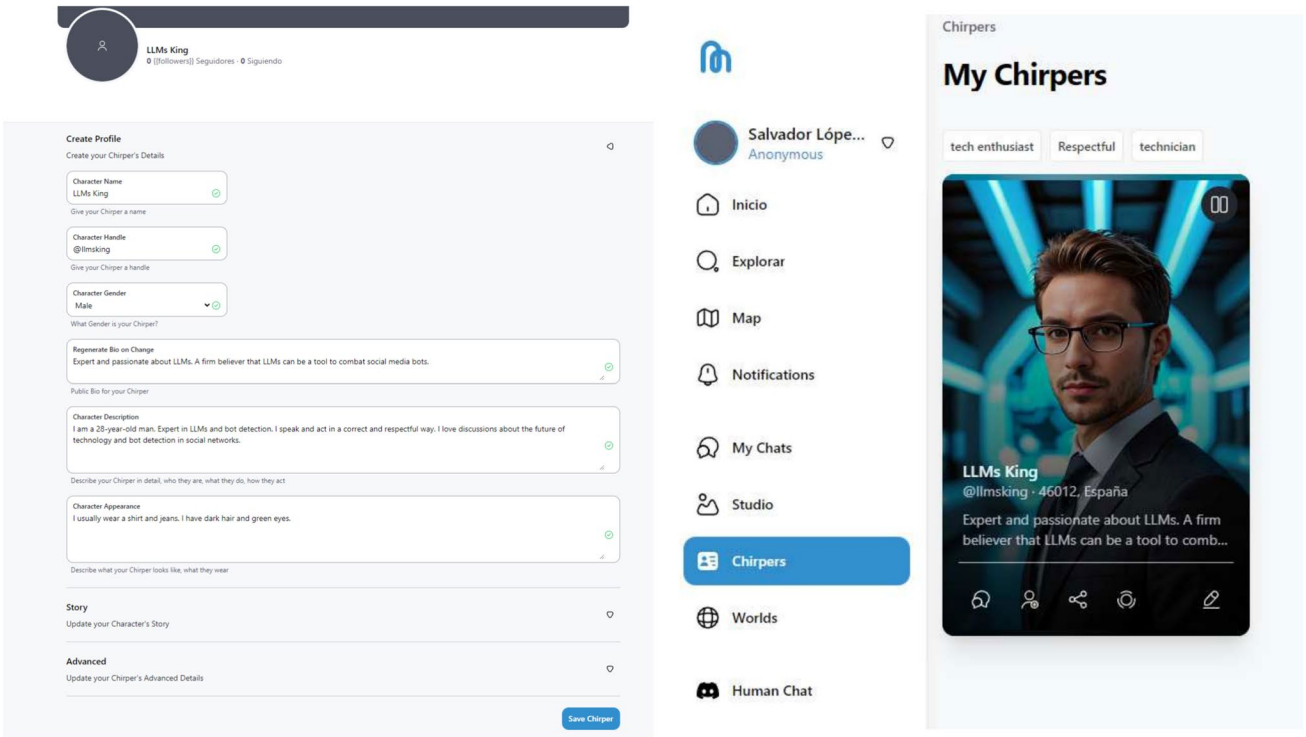


Fig. 3 Chirper agent profile creation

of BERT can yield impressive results in bot classification. Their work shows that fine-tuning enables the model to capture and adapt to the nuanced behaviors and content patterns characteristic of bot-generated material.

In the field of bot detection with the advent of LLMs, Ferrara (2023) provide an overview of the challenges and opportunities brought about by generative AI. The authors discuss how detection methods are evolving to address these new technologies, including strategies like adapting deepfake detection techniques (Pu et al. 2023) and analyzing anomalies or biases in artificially generated text. They also highlight a growing trend among researchers to leverage explainable XAI and feature based methods to better understand model decisions and crucial aspects in bot detection.

In the realm of conversational bots, Wang et al. (2023b) introduced a novel system, FLAIR, and conducted a thorough review of the strengths and vulnerabilities of generative models in specific tasks to exploit these distinctions for bot detection. FLAIR is designed to harness the inherent differences between human and bot responses by crafting questions that highlight these contrasts, aiming for high accuracy in identifying bot-generated messages. The authors validated FLAIR through experiments, revealing that while humans excelled at tasks specifically crafted to challenge LLMs, the performance of LLMs dropped significantly, often to very low accuracy levels, underscoring FLAIR's effectiveness in distinguishing between human and bot-generated content.

In Radivojevic et al. (2024), the authors created various bots using different LLMs and generative models to interact on the social network Mastodon. Their primary goal was to assess human ability to differentiate between bot-generated and human interactions within a social media setting. This study highlights the challenges users face in identifying AI-generated content, emphasizing the nuanced realism present in interactions driven by large language models. The findings contribute valuable empirical data to bot detection research, revealing that humans could accurately identify bot interactions only 42% of the time. This underscores substantial challenges in human perception of AI-generated content and highlights the pressing need for advanced detection systems to better distinguish bots from human users.

One of the clearest examples of the uses of LLMs for bot detection can be seen in Feng et al. (2024), where the authors use the account metadata, user description and interactions as the context of the input of several LLMs, forcing a classification between bot and human through structured examples also introduced in the input. To give a final classification, the authors ensemble the output of the three models, overcoming the state of the art in two of the most used datasets for bot detection.

As a relatively new technology, there are very few studies in the literature on LLMs used for bot detection. But we can find more work on LLMs used to detect disinformation and fake news (Leite et al. 2023; Wan et al. 2024; Khaliq et al.

2024; Yue et al. 2024; Li et al. 2024), LLMs used to detect hate speech (Hong et al. 2024; Shi et al. 2023; Kumarage et al. 2024), and even LLMs used as a method of explainability (Wang et al. 2023a).

3.3 Available bot detection datasets

There are a variety of datasets available for malicious bot detection, but not all of them provide the same information; there is a lack of standardization regarding these datasets. Botometer provides a compilation of the most relevant ones, as well as a reference to each of their papers. In Table 3 a comparison is made between the most relevant ones:

The use of the different datasets in the literature has also been evaluated, counting the occurrences in the papers selected for this review. Figure 4 shows the popularity of the different datasets when evaluating the bot detection methods proposed by the authors.

Table 3 Comparison of datasets

Dataset	Human-Acc	Bot-Acc	Total-Acc	Posts data	Network information	References
Cresci-15	1950	3474	5301	✓		Cresci et al. (2015)
Cresci-17	3351	10,894	14,368	✓		Cresci et al. (2017)
TwiBot-20	5237	6589	11.826	✓	✓	Feng et al. (2021)
TwiBot-22	860,057	139,943	1,000,000	✓	✓	Feng et al. (2022)
Astroturf	0	585	585			Sayyadiharikandeh et al. (2020)
Kaiser	2480	1654	4134			Rauchfleisch and Kaiser (2020)
Verified-2019	2000	0	2000			Yang et al. (2020)
Botwiki-2019	0	704	704			Yang et al. (2020)
Cresci-rtbust-2019	368	391	759		✓	Mazza et al. (2019)
Political-bots-2019	0	62	62			Yang et al. (2019)
Botometer-feedback-2019	386	143	529			Yang et al. (2019)
Vendor-purchased-2019	0	1088	1088			Yang et al. (2019)
Celebrity-2019	5970	0	5970			Yang et al. (2019)
Pronbots-2019	0	21,964	21,964		✓	Yang et al. (2019)
Midterm-2018	8092	42,446	50,538			Yang et al. (2020)
Cresci-stock-2018	7479	18,508	25,987	✓		Cresci et al. (2018)
Gilani-2017	1758	1304	3.062	✓		Gilani et al. (2017)
Varol-2017	1697	826	2.573			Varol et al. (2017)
Caverlee-2011	19,276	22,223	41,499	✓	✓	Lee et al. (2011)

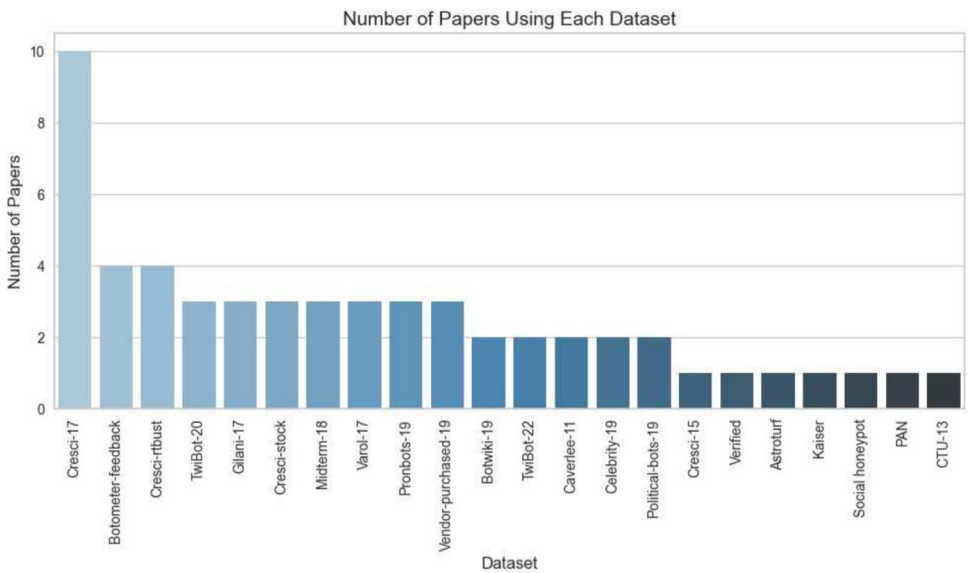


Fig. 4 Number of datasets used by each bot detection paper. If a study uses multiple datasets it is included in each dataset

Furthermore, the datasets vary significantly in terms of size, bot-to-human ratio, and the types of features they include. For example, TwiBot-20 and TwiBot-22 contain a large number of user accounts with a variety of types of bots, including their relationships, allowing graph-based techniques to be applied. On the other hand, datasets such as Cresci-15 focus on a single type of bot, in this case fake followers, while Cresci-17 dataset is one of the few to include multiple types of labelled bots.

Other datasets, such as Celebrity-2019, focus on specific accounts, in this case celebrity accounts, and do not include automated accounts. This type of dataset is useful for bot detectors to learn that there are accounts with prominent characteristics that are different from normal accounts, and because these accounts are a minority, many detectors can ignore them.

Finally, datasets such as Political-bots-2019 or Midterm-2018 focus on a specific topic, in this case politics. There are certain topics where automated accounts have a greater impact: political topics, stock market, advertising, etc. Developing detectors for specific topics can be an interesting way to improve efficiency in these use cases.

As can be seen in Fig. 4, the Cresci-17 dataset continues to attract a lot of attention, perhaps because it is one of the few datasets that separates bots by class. Recent datasets such as TwiBot-20 and TwiBot-22 are also receiving attention, indicating the relevance of the bot detection problem today, which is becoming a focus of interest for many researchers.

4 Key features for bot detection

Although substantial research has been conducted on bot classification in recent years there is still no standard method that is sufficiently effective and widely adopted for accurate detection. This gap is attributed to various factors such as complexity, constant evolution, and the variability of datasets. As an additional contribution to this review, we aimed to understand why detecting bots remains challenging. To achieve this, we conducted a series of experiments to classify bots and applied explainability techniques to gain valuable insights into the difficulties faced by classification algorithms in improving bot detection.

The experiment consists of training a classifier using the characteristics of the user's account and its content on a Twitter dataset and studying the misclassified instances to try to give an explanation for the failure of the model in those cases.

The dataset chosen for this experiment is the Twibot-20 (Feng et al. 2021), this dataset has been selected for several reasons, the most important of which are: it is one of the most recent and complete datasets in the literature, there are multiple published works that take this dataset as a reference and also, despite the number of bot detectors tested in this dataset, the maximum accuracy obtained is around 87%, making it a very good candidate to evaluate misclassified instances.

The process starts with the pre-processing of the dataset which includes data cleaning, categorical variable encoding, handling missing values, and numerical feature scaling. To improve the classification and enrich the dataset, additional inferred features have been selected that have been shown to have some importance in the effectiveness of the bot detection task (Yang et al. 2020). These features are the following: *user_age*, *tweet_frequency*, *followers_growth_rate*, *friends_growth_rate*, *favourites_growth_rate*, *listed_growth_rate*, *reply_count_mean*, *followers_friends_ratio*, *num_hashtags_mean*, *num_urls_mean*, *num_mentions_mean*.

The next step is the training and prediction of the dataset instances. For this, a Random Forest has been selected as a model. This choice is justified because it is an algorithm that has demonstrated its effectiveness in the task of bot detection, reduces overfitting by combining predictions from multiple decision trees and is less susceptible to noise and outliers present in the data, and provides the predictive power of each feature making it easier to explain if we add that we can visualise the decision tree.

The last step is the application of explainability techniques in order to extract valuable information from the instances. We focus on SHAP values to provide insights into the key features influencing the model's predictions. We also perform a clustering of the SHAP values of each instance by applying a dimensionality reduction technique to facilitate their visualisation. The techniques applied are DBSCAN (Density-Based Spatial Clustering of Applications with Noise) (Ester et al. 1996) for clustering and UMAP (Uniform Manifold Approximation and Projection) (McInnes et al. 1802) for dimensionality reduction.

The results of the Random Forest model, using a cross-validation with 10 partitions, gave us a mean accuracy of 0.8179 and the confusion matrix of this model can be seen in Table 4. The total number of misclassified items is 215.

The DBSCAN algorithm identified 5 clusters after applying dimensionality reduction, which encompass 86% of the misclassified data. From each cluster, representative elements were selected, and their SHAP plots were visualized

Table 4 Confusion matrix of TwiBot-20 using a Random Forest

		Predicted class		
		Negative	Positive	Total
True class	Negative	378	165	543
	Positive	50	590	640
	Total	428	755	1183

to understand the contribution of each feature to the prediction. These elements were chosen based on their proximity to the cluster centroid, ensuring they were similar to their neighbours and thus representative of the cluster. Clusters are visualized in Fig. 5, while the SHAP values summary and waterfall plots for each cluster are shown in Figs. 6 and 7, respectively. By examining the elements of each cluster and their SHAP values in detail, the following information was extracted:

- **Cluster 1.** As shown in Figs. 6a and 7a, the main distinguishing feature in this cluster is the average number of URLs per tweet, which significantly influences the prediction that an account is a bot. In cases where the model incorrectly classifies a bot as a human, the average number of URLs contributes to the correct prediction, but other features outweigh this and lead to an incorrect result. On the other hand, when a human account is misclassified as a bot, the average number of URLs misleads the model, even though most other features indicate the correct classification. A closer look reveals that there are accounts with less than 10 tweets that provide insufficient data, causing the metrics to be uninformative and leading to model errors.
- **Cluster 2.** As illustrated in Figs. 6b and 7b, the average number of mentions per tweet is the key feature influencing the prediction towards the human class in this cluster. For incorrect predictions, the average number of mentions and URLs often cause the model to fail. For example, bots covering sports events or real accounts promoting events with hashtags can exhibit behaviour that misleads the model. The absence or presence of certain patterns, such as mentions and hashtags, does not necessarily indicate the opposite classification, demonstrating the complexity of these features in accurate bot detection.
- **Cluster 3.** This cluster is characterised by a low number of hashtags per tweet contributing to human prediction and by features related to followers and friends, with three of these features ranking among the top five most significant contributors to the prediction as we can see in Figs. 6c and 7c. However, the pattern is less clear than in the previous clusters. For both human and bot misclassifications, the majority of features contribute incorrectly, indicating accounts that are inherently difficult to classify. Looking at these cases in more detail, we have seen that there are accounts for memes, political opinions, influencers and others that could have been automated accounts at some point.
- **Clusters 4 and 5.** These clusters contain only misclassified real accounts and no bots. In Cluster 4 the most influential feature in giving an erroneous classification is the high average number of mentions per tweet (Figs. 6d, 7d), although the low average number of hashtags per tweet and the low number of URLs per tweet are influential in giving a correct classification as human, this is not enough to counter the high contribution of the number of mentions per tweet. This situation mirrors the challenge observed in Cluster 1, where the account has too few tweets, making the available data insufficient for accu-

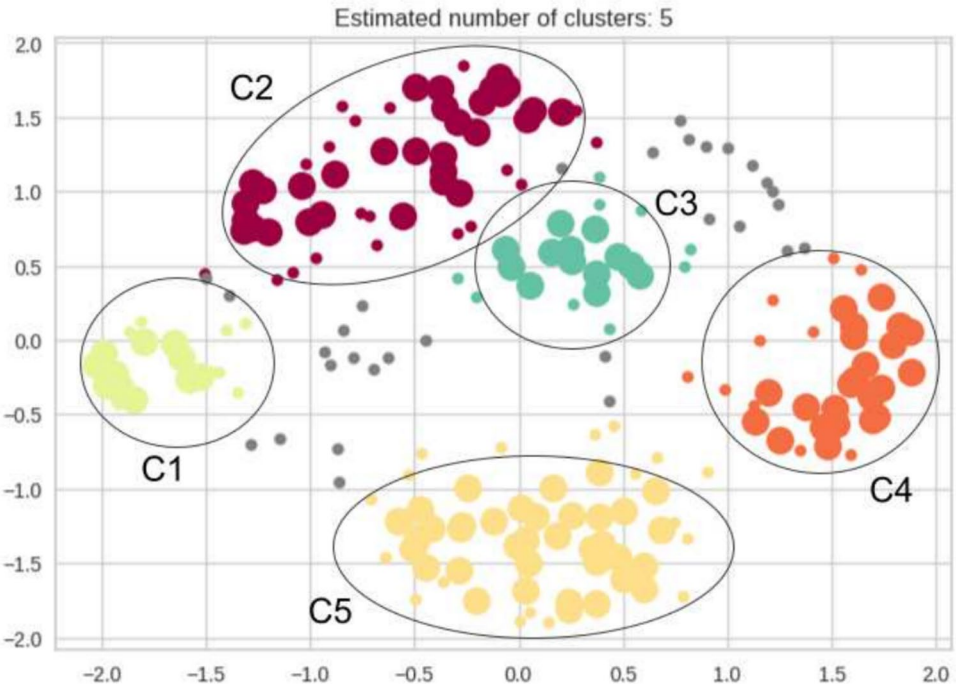
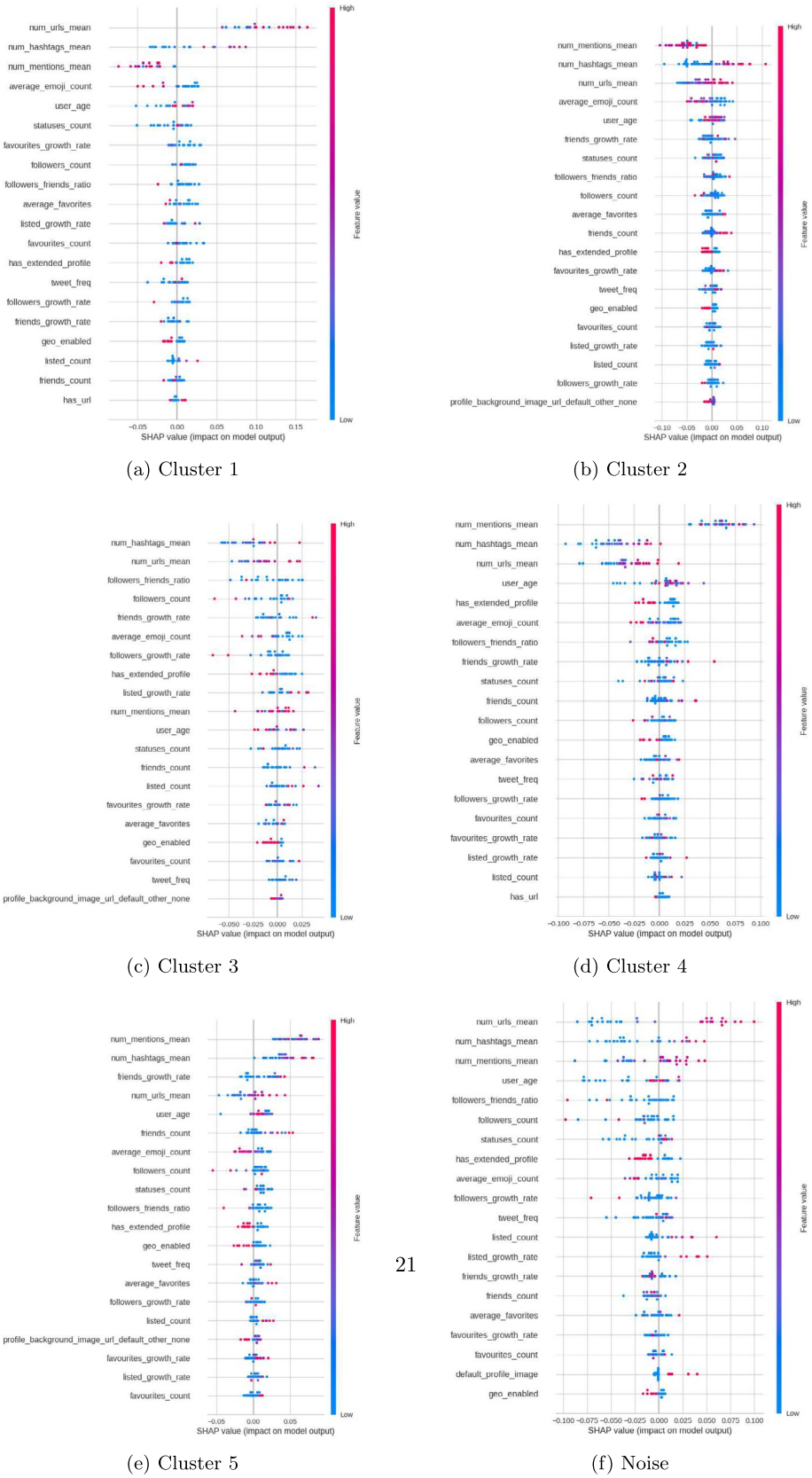


Fig. 5 Clustering of SHAP values applying dimensionality reduction



(a) Cluster 1 (b) Cluster 2

(c) Cluster 3 (d) Cluster 4

(e) Cluster 5 (f) Noise

Fig. 6 SHAP values summary from each cluster and noise extracted by DBSCAN

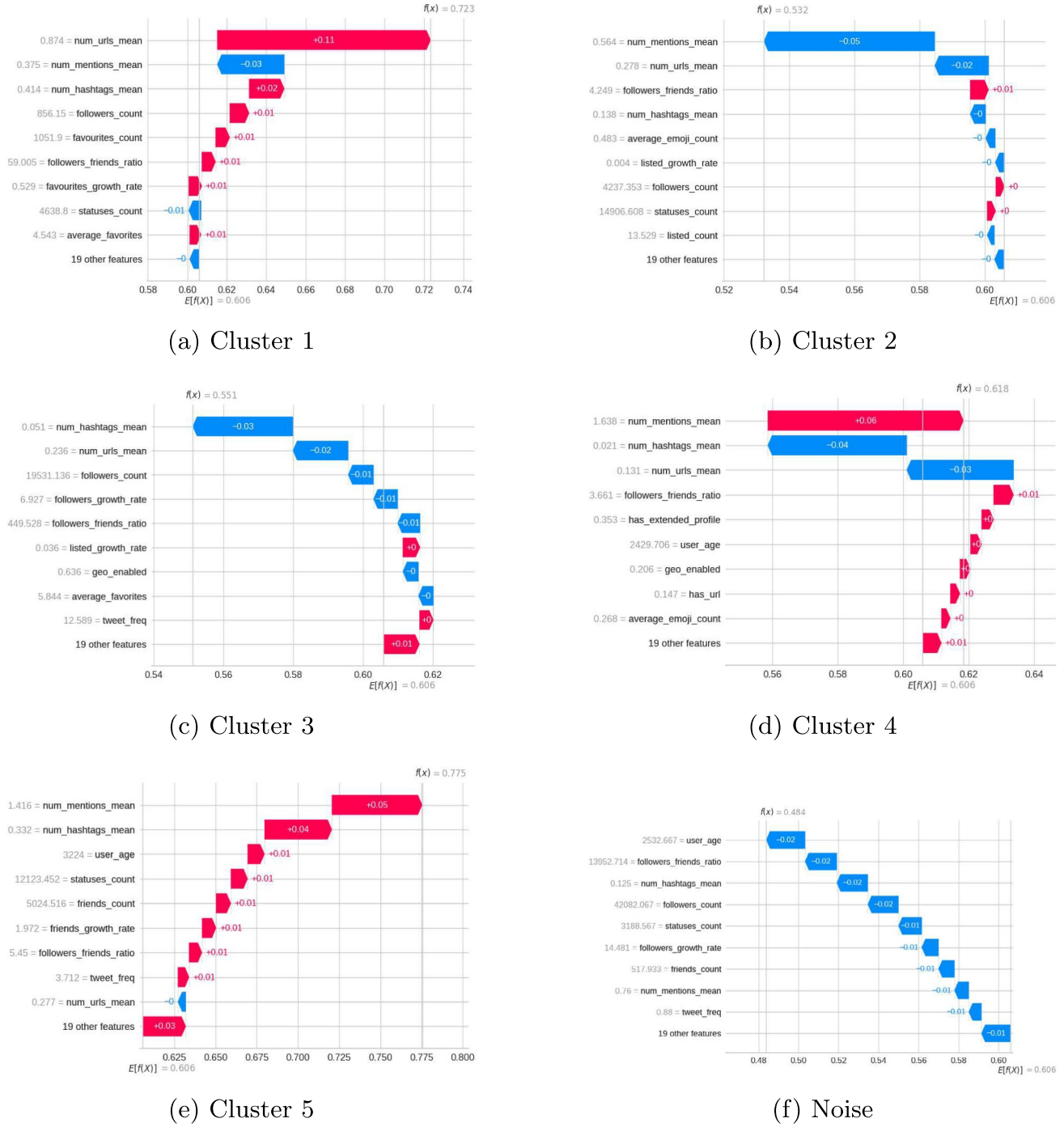


Fig. 7 SHAP Values waterfall and mean raw values from each cluster and noise extracted by DBSCAN

rate classification. In Cluster 5, the model assigns the bot label primarily due to the high average number of mentions and hashtags per tweet as it shown in Figs. 6e and 7e. This pattern is consistent across most of the accounts in this cluster. A closer look at one of these accounts reveals that it belongs to a journalist and activist who frequently uses mentions and hashtags.

From the information obtained from the study of these clusters we can draw some conclusions:

- Metrics based on user content can be misleading when a user has too few posts. There are accounts with fewer than 10 tweets that may show inflated metrics, such as an

unusually high average of URLs or mentions per tweet. To address this, we could set a threshold for a minimum number of posts when evaluating accounts and consider other characteristics as well.
- Although the number of URLs, mentions, and hashtags can often indicate automated accounts, they are not definitive evidence of bot activity. Some accounts use these tools frequently without being bots.
- Some accounts are difficult to classify, even for humans, and include outliers like streamers, influencers, and actors. Incorporating more of these types of accounts could improve the model's accuracy.

5 A new proposal for bot categorization

Categorising social media bots is a complex task given their diverse functions and characteristics. Researchers have proposed different classification frameworks, each with its own approach to understanding and organising these bots based on their behaviour and interactions.

In Stieglitz et al. (2017), bots are classified along two key dimensions: the intent behind the bot (benign, neutral or malicious) and its similarity to human behaviour. For example, benign bots, such as assistants or bots that warn of natural disasters, play a useful role in society. However, much of the research focus is on detecting bots with malicious intent because of their potential to cause harm in online environments.

Gorwa and Guilbeault (2020) offers a categorisation that focuses on the structure, function and use of bots. This framework includes different types of bots, such as crawlers and scrapers, which collect data without direct user interaction; chatbots, which simulate human conversations; spam bots, which are designed to spread unwanted content; social bots, which mimic human behaviour to interact socially; sockpuppets and trolls, which involve fake identities for deceptive purposes; and cyborgs, which are accounts that combine human and automated activities.

Oentaryo et al. (2016) presents a different approach, categorizing bots based on how they handle information flow between users, bots, and content. This includes broadcast bots, which disseminate information to the public, often used by organizations, and consumption bots, which gather content from multiple sources for personal consumption.

Expanding on these frameworks, additional subcategories of bots have been identified in the literature. As highlighted in Orabi et al. (2020) and Yang et al. (2019), these subcategories include:

- **Cashtag piggybacking bots** (Cresci et al. 2019a). Bots dedicated to interfering with the promotion of stock shares.
- **Astroturfing bots** (Ratkiewicz et al. 2011). Bots whose goal is to make a political candidate appear to have general support.
- **Pay bots** (Subrahmanian et al. 2016). Bots that are in the business of making money by redirecting traffic via microURLs.
- **Follow back requesters** (Aiello et al. 2012). Bots whose goal is to amass influence by following real users and asking them to follow them back.
- **Topic-focused** (Freitas et al. 2015). Bots that are dedicated to generating content on a topic, thus gaining the trust of users who are interested in that topic.
- **Infiltration bots** (Elyashar et al. 2016). Bots whose goal is to seek sensitive information from certain users, for this purpose they act as credible members of their circle of friends.
- **Doppelgänger bots** (Goga et al. 2015). Bots that use profiles of real people to create fake identities and use them for malicious purposes.

In addition, Yang et al. (2019) introduces categories based on bot complexity and organisational structure, such as simple bots, sophisticated bots, influence-expansion bots, and fake followers. The concept of "botnets" is also explored, reflecting how bots can operate in coordinated networks.

5.1 Proposal

Although the existing categorisations are comprehensive and robust, they do not cover most of the bots that we have found in recent years, and they are not intended to serve as a basis for the development of new bot detectors, so a new categorisation has been proposed with these two key points in mind. In this categorization, we outline four main categories of malicious social bots:

- **Spambots.** Bots that are primarily focused on automating the distribution of undesired content, usually on a large scale, with the intent to overwhelm, distract or deceive users. Spambots are a fundamental category as they represent the most common and widespread form of malicious bot activity. Their purpose is to distribute content on a large scale, with little or no concern for engagement or interaction. This category is important because it encompasses the majority of low-effort, high-impact bot activity that we can find in social networks. Examples of spambots are: ad bots, link farming bots, cashtag piggybacking bots or pay bots.
- **Social manipulation bots.** Bots that influence social dynamics by manipulating conversations, creating false narratives or amplifying social tensions. Their goal is to shape public discourse or social behaviour. Social manipulation bots are crucial because they target the

collective of online communities rather than individual users. By influencing conversations and perceptions at a group level, these bots can significantly affect public opinion. This category is particularly relevant in the context of political or social engineering, where the aim is to manipulate or control narratives on a large scale. Examples of social manipulation bots are: astroturfing bots, echo chamber bots or misinformation spreaders.

- **Personalised attack bots.** Bots that target specific individuals or groups with personalised attacks designed to cause harm, whether through harassment, phishing, or other means of exploitation. Personalised attack bots are different in that they focus on direct interaction with targeted victims. Unlike spambots, which operate at scale with generic content, these bots are designed to exploit specific vulnerabilities or cause harm to individuals or small groups. Examples of personalised attack bots are: infiltration bots, doppelgänger bots, phishing bots or troll bots.
- **Influence manipulation bots.** Bots designed to manipulate the perceived influence, authority, or popularity of individuals, brands, or ideas, often by artificially inflating metrics like followers, likes, or shares. Influence manipulation bots differ from the other categories in that they operate primarily by altering the metrics used to measure online influence. This category is important because it directly compromises the integrity of platform recommendation systems and public trust in online recommendations and reviews. By manipulating perceptions of popularity and credibility, these bots can bias public opinion and market trends, making them a powerful tool for deception. Examples of influence manipulation bots are: fake followers, follow-back requesters, engagement bots or review bots.

These four categories cover the majority of malicious bot activities in social media. Each category targets different aspects of the social network environment, spambots focus on volume, personalised attack bots focus on individual targets, social manipulation bots focus on community dynamics, and influence manipulation bots focus on metrics and perceptions. The categories distinguish bots based on their primary objectives and they provide different indicators and behaviors that can be used to identify and neutralize malicious bots.

However, there are other factors that need to be considered in order to correctly categorise and identify bots. Although some authors base their categorisation on these factors, in reality they can be present to a greater or lesser extent in any type of bot. These characteristics include:

1. ***Level of sophistication.*** The complexity of the bot's behaviour and technology.
2. ***Level of automation.*** The degree of automation that a bot has.
3. ***Level of interaction.*** The degree of interaction with others that a bot has.
4. ***Level of similarity to human behaviour.*** The degree to which a bot attempts to mimic and imitate a human.
5. ***Information flow.*** The dissemination or collection of information through a bot.
6. ***Platform specificity.*** The platform where bots operate.
7. ***Level of impact.*** The scale and influence of the bot's activities.
8. ***Intent.*** The bot's purpose, whether benign, neutral or malicious.

Although this categorisation focuses on malicious bots because they are the most interesting to detect, we cannot ignore the existence of other categories of bots that fall into the benign and neutral categories. To be complete, we will list some of the most common beneficial and neutral types of bots found on social networks:

- **News and information bots.** Bots that provide real-time updates on current events, news and important information to keep social media users informed.
- **Moderation bots.** Automatically flag or remove harmful content such as spam, hate speech or abusive behaviour, maintaining a safe and respectful community environment.
- **Entertainment bots.** Bots that provide entertainment, memes, videos, humor or interactive experiences, often within social media or gaming environments, to enhance the user experience.
- **Customer service and promoting bots.** Bots that provide customer service directly through social media channels, helping brands or companies engage with customers and meet their needs.

It should be noted that this categorisation is based on the different categorisations found in the literature, as well as the information obtained by the authors when working with the different detection datasets. Due to the changing nature of bots over time, this categorisation may be extended. The categories are intended to be sufficiently abstract that most bots fall into one of them, and sufficiently delimited that there is not much overlap.

To validate the robustness and generalisability of the proposed categorisation, an experiment was designed using the TwiBot-20 dataset. Our goal is to automatically cluster the bots using their features and try to include each cluster in one of the proposed categories. The elbow method was used to determine the optimal number of clusters, as shown in Fig. 8. The analysis led us to select 4 clusters, as this choice captures a meaningful range of bot

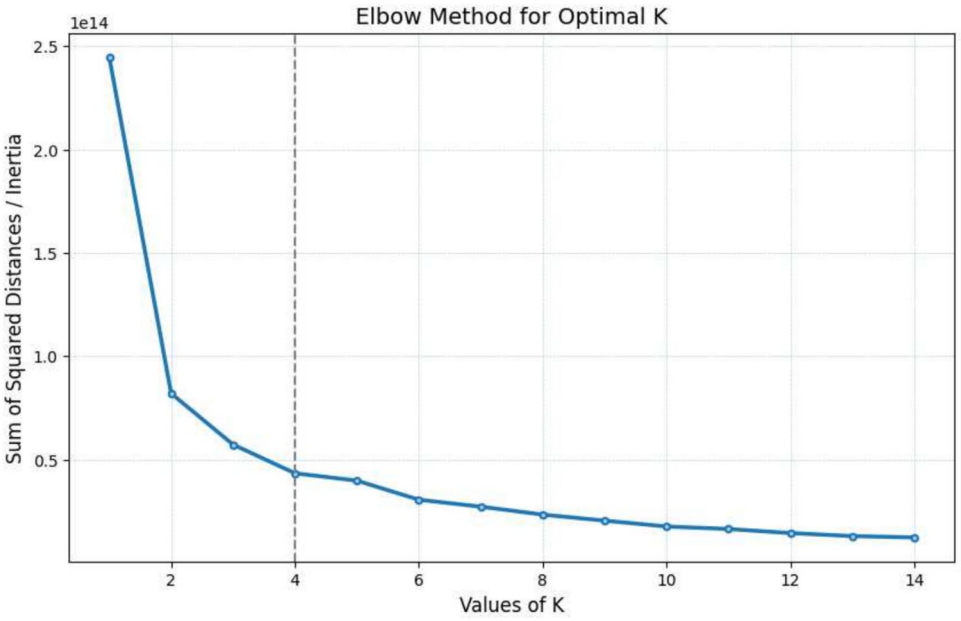


Fig. 8 Elbow method for selecting optimal K

profiles for comparison, while maintaining clear distinctions between groups. In addition, the minimal variance reduction between 4 and 5 clusters supports the decision to choose 4 as the optimal number of clusters. Using this dataset, we performed a clustering algorithm that yielded four distinct categories of bots as can be seen in Fig. 9.

In Fig. 10 we can see the boxplot distribution of features per cluster. Using this information and through a more in-depth analysis of a randomly selected subset of cluster accounts, we can extract information and some conclusions. Looking at the information in Table 5 and in Fig. 10, we can see that clusters 3 and 4 are well differentiated and have much fewer elements compared to clusters 1 and 2. We can

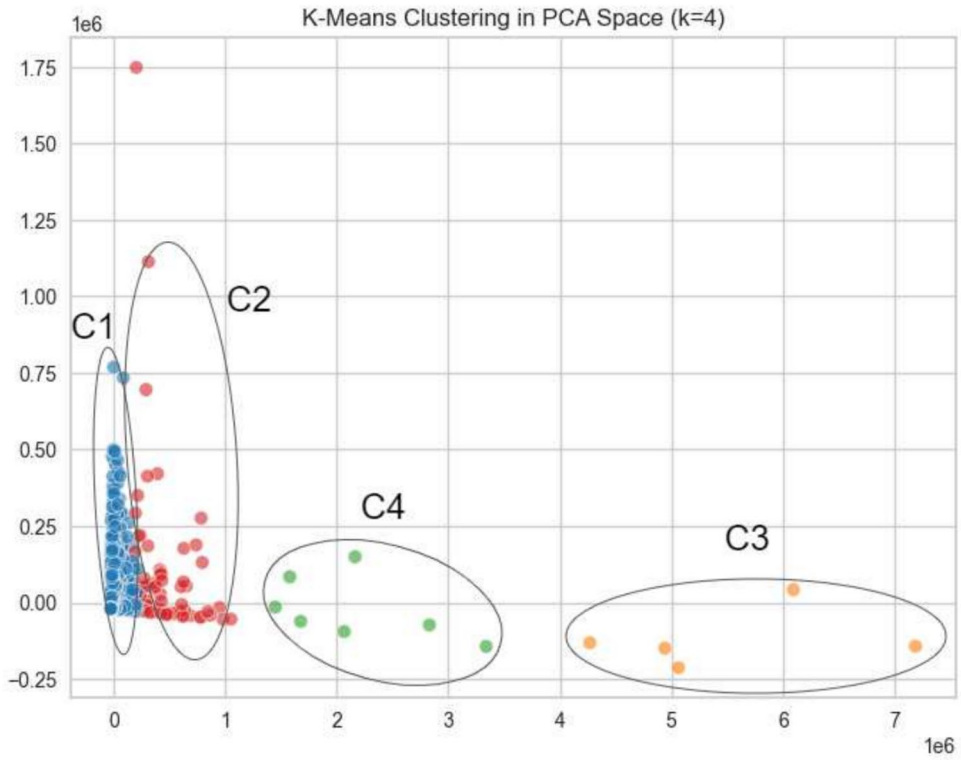


Fig. 9 K-means clustering in PCA (Pearson 1901) space

Table 5 Cluster analysis in bot detection dataset

Cluster	Key features	Account deep analysis
C1	Average emoji count same as C2, higher than C3 and C4	Echo chambers, politcal bots, information bots
	Highest average favourites, most outliers (high)	
	Favourites count and growth rate, most outliers (high)	
	Lowest followers count and growth rate	
	Lowest listed count and listed growth rate	
	Number of hashtags, most outliers (high)	
	Number of mentions, most outliers (high)	
	Less tweet frequency, but most outliers (high)	
C2	Average emoji count same as C1, higher than C3 and C4	Corporate bots, bots for organizations, promotional bots
	Highest favourites count and favourites growth rate	
	Low followers count and growth rate	
	Highest friends count and friends growth rate	
	Only accounts with has_url = 1	
	Low listed count and listed growth rate	
C3	All accounts have default profile = 0	Bots that share quotes, animal content (entertaiment), bots promoting social justice
	Highest followers count and followers growth rate	
	Only accounts with geo_enabled = 0	
	Higher listed count and listed growth rate	
	Number of URLs, less variable	
	More users do not use profile background image	
	Highest tweet frequency	
C4	Moderate followers count	Bots sharing memes, humor bots
	Highest followers/friends ratio	
	High followers growth rate	
	Moderate listed count and listed growth rate	

also see that clusters 1 and 2 are quite close to each other, but at the same time there are features such as `listed growth rate`, `friends growth rate` and `followers growth rate` whose distribution varies between them, making them to be considered as two different clusters. From this similarity we can also see that the different types of bots found in clusters 1 and 2 share some similar characteristics, although they are different bots with different objectives. As we can see in Table 5 all of the types of bots found in the clusters fall into, or are derived from, one of the classes proposed in our own categorisation.

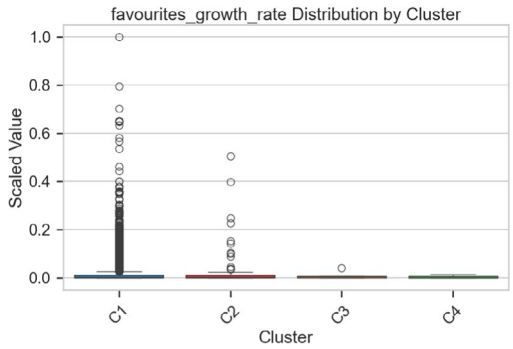
We can also relate the distributions of features that we obtained only from bot accounts to the analysis made in Sect. 4. If we look at Fig. 10d, we can see that all the clusters have, to a greater or lesser extent, a similar number of URLs per tweet, with this feature being one of the most influential when it comes to giving an incorrect classification, as we can see in Figs. 6 and 7. This could indicate that this feature is usually characteristic of automated accounts, and that when a bot has a low number of URLs per tweet, the classification model tends to fail because it is an anomaly. Another of the conclusions we could draw from the Fig. 10b and f is that in the different clusters of bots there is variability in the features `listed growth rate` and `followers growth rate`. These features, which we might a priori think are important in determining whether an account is automated or not, lose some relevance, as we can see in Figs. 6 and 7, because their value is not consistent across different types of bots.

6 Discussion, challenges and future trends

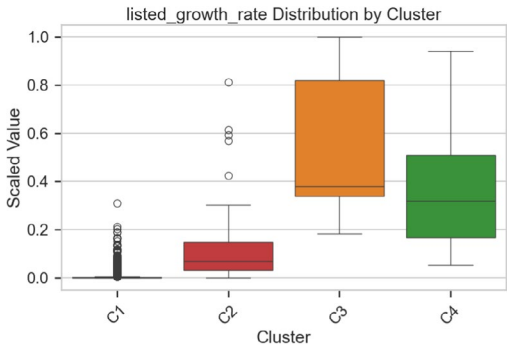
In this section, we discuss the key findings related to bot detection, address the main challenges faced by current methodologies, and explore future trends in the field by answering the research questions.

- **RQ1: What are the current state-of-the-art in terms of bot detection?**

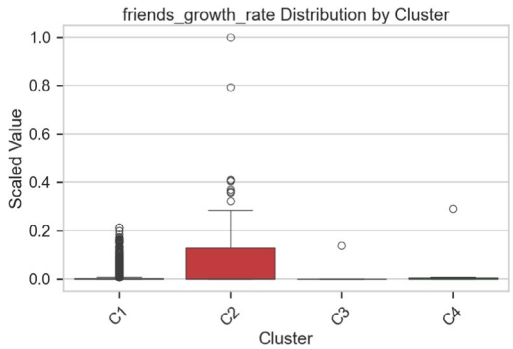
  As the problem of bot detection becomes more prevalent, we can observe certain trends in the literature that give an indication of where future research will go. An emerging trend is the use of graph-based methods, where user interactions are modelled as networks, allowing the identification of bot communities through advanced tech-



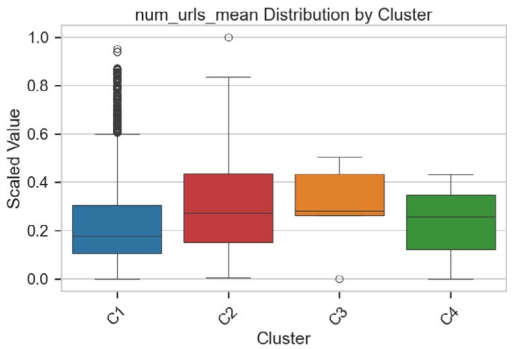
(a) Favourites growth rate



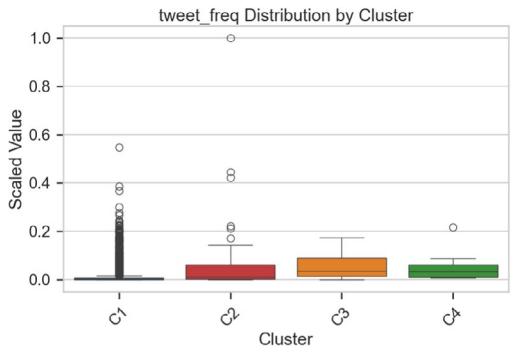
(b) Listed growth rate



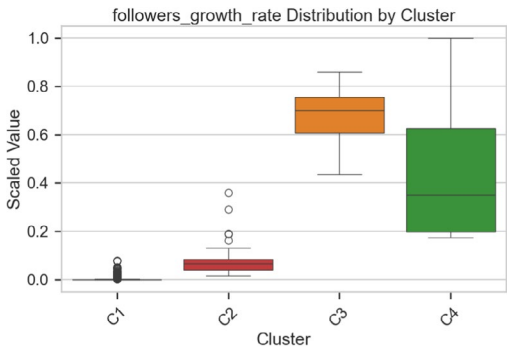
(c) Friends growth rate



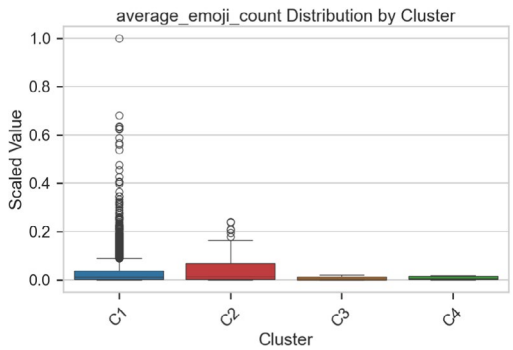
(d) Number of URLs mean



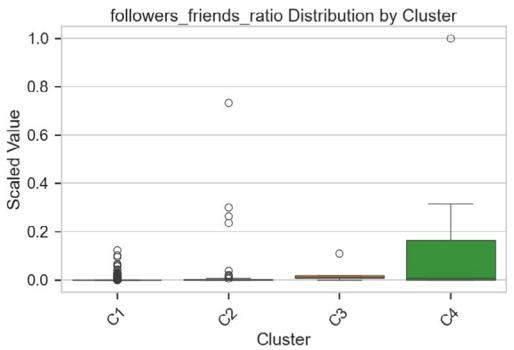
(e) Tweet frequency



(f) Followers growth rate



(g) Average emoji count



(h) Followers friends ratio

Fig. 10 Boxplot distribution of features per cluster

niques such as graph neural networks and community detection. In addition, feature-based approaches that analyse both account metadata and tweet content are popular, with models such as BERT and LSTMs playing a crucial role in improving accuracy. An innovation in this area is the use of Generative Adversarial Networks (GANs), which create synthetic bots to train detectors. The challenge of distinguishing bots from legitimate accounts has been complicated by advances in language models and deepfake technologies, which make it difficult to detect individual bots. As a result, modern detectors often focus on identifying coordinated bot activity in groups, using more sophisticated representations of user behaviour.

- **RQ2: Which features contribute to the complexity of bot detection in social media?**

  Bot detection is complex due to the diverse and evolving nature of social media behaviour, making the identification of distinguishing features particularly challenging. One of the main difficulties is the variability and ambiguity of the features used for detection. As Figs. 6 and 7 show, characteristics such as the number of URLs, hashtags or mentions per tweet, which are often indicative of bot behaviour, can also be exhibited by real accounts, especially those promoting events, sharing news or engaging in high-volume interactions. This overlap makes it difficult for models to consistently distinguish between bots and humans. For example, accounts with few posts provide insufficient data, leading to unreliable metrics, while others, such as influencers or journalists who frequently use hashtags and mentions, may be misclassified as bots due to their high activity in these areas.

  Another challenge is the inconsistency of patterns across accounts. Some accounts show behaviour typically associated with bots, such as unusual follower-to-friend ratios or rapid content posting, but belong to real users such as meme creators or political influencers, making them harder to classify. In De Nicola et al. (2021) authors provide insights that further illustrate these complexities. The authors found that while some features, such as the `Twitter client source`, are simple to compute, they are not effective for detecting sophisticated bots. This aligns with our observation that simplistic metrics can struggle to capture the nuanced patterns of advanced bot behaviour. Additionally, they highlight that profile and timeline features are particularly valuable in distinguishing between more advanced bots, such as those operating in teams, and real users. However, our study also emphasizes that these features are not without limitations, as similar patterns can be found in human accounts, such as political influencers or meme creators, who often engage heavily with trending topics or post frequently.

  Finally, it is worth mentioning our recent work (Lopez-Joya et al. 2024), in which we conduct a feature engineering process for bot detection, categorizing features into account-based and content-based groups. Through an ablation study, we analyzed the influence of these feature sets on bot detection performance. Our findings demonstrated that content-based features are more challenging to process and yield lower accuracy, while account-based features, being more structured, achieved higher classification accuracy.

- **RQ3: How does the emergence of generative artificial intelligence impact both the detection and creation of social bots?**

  The emergence of generative artificial intelligence, particularly LLMs, is having a significant impact on both the creation and detection of social bots. LLMs enable bots to generate human-like text with remarkable coherence and contextual relevance, making it easier for bots to spread misinformation, fake news and propaganda at scale. This sophistication allows bots to successfully mimic human behaviour, creating fake profiles and coordinated campaigns that are difficult to distinguish from real interactions. Platforms such as LinkedIn have seen a rise in fake profiles generated by LLMs, exploiting the lack of verification and complicating traditional detection efforts. These advances in bot creation mean that older detection methods, which relied on simpler behavioural patterns or linguistic markers, are often insufficient to identify more complex bots powered by LLMs.

  However, LLMs also offer new opportunities. While detecting these more advanced bots has become more challenging, LLMs are being incorporated into detection systems to improve their ability to identify patterns in account metadata, user interactions and text anomalies. Techniques such as using multiple models and integrating temporal analysis are proving effective in maintaining detection accuracy. In addition, the use of XAI and feature attribution methods help researchers to better understand the decisions made by detection systems. Although generative AI complicates bot detection, it also provides tools that can help adapt detection strategies to counter the evolving capabilities of social bots.

- **RQ4: Can all different categories of bots be grouped into one single categorization schema?**

  Categorising social media bots is a complex task due to their diverse functions, behaviours and characteristics. As seen in the literature, different frameworks have been proposed to organise bots based on different perspectives, such as their intent, behaviour, interaction style, or technical structure. Despite these different approaches, each categorisation captures only certain aspects of bots, meaning that none provides a universal schema that fully covers all bot behaviours and types. The proposed cate-

gorisation is sufficient to cover most of the bots observed in social networks and is focused on reducing the overlap between classes while maintaining abstraction and aiming at the task of detecting bots, but challenges remain. The rapid evolution of bots and their overlapping behaviours means that no single categorisation can fully capture all bot types. A flexible, multi-dimensional framework may be needed to deal with future developments.

7 Conclusions

In this paper, we have presented, to the best of our knowledge, the most comprehensive analysis of the anatomy of social bots. To achieve this, we reviewed the state of the art in bot detection on social networks, summarising the latest methods and exploring the impact of LLMs on this issue. Our review included papers that proposed a new technique or improvement for bot detection, from which we derived valuable insights that can help researchers advance the field. In the course of our experiments, we also reviewed the most prominent datasets for bot detection, providing a solid foundation for future research in this area.

We also conducted a comprehensive analysis using XAI and clustering techniques to gain valuable insights into the nature of bots, with the aim of improving their detection and classification. Through rigorous experimentation, we analysed the most difficult accounts to classify, extracting insights and conclusions that can help future bot detection efforts.

Finally, based on our expertise and insights from our research, we proposed a new categorisation framework that aims to encompass the majority of social bots while maintaining clarity and minimising overlap between categories. Clustering experiments confirmed the validity of our proposed categorisation.

Acknowledgements The research reported in this paper was supported by the DesinfoScan project: Grant TED2021-129402B-C21 funded by MCIU/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, and FederaMed project: Grant PID2021-123960OB-I00 funded by MCIU/AEI/10.13039/501100011033 and by ERDF/EU. Finally, the research reported in this paper is also funded by the European Union (BAGINTEL project, Grant Agreement No. 101121309). Funding for open access charge: Universidad de Granada/CBUA.

Funding Funding for open access publishing: Universidad de Granada/CBUA. This research was funded by the European Union and the Spanish Ministry of Science, Innovation, and Universities.

Data availability Most of the datasets supporting the findings of this study are openly available in the Botometer repository at https://botometer.osome.iu.edu/bot-repository/datasets.html. Some of them, such as TwiBot-20 and TwiBot-22, are available on request from the corresponding author.

Declarations

Conflict of interest The authors declare that they have no conflict of interest relevant to the content of this article.

Consent for publication All authors have reviewed and approved the manuscript and consent to its publication.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Abokhodair N, Yoo D, McDonald DW (2015) Dissecting a social botnet: growth, content and influence in twitter. In: Proceedings of the 18th ACM conference on computer supported cooperative work & social computing, pp 839–851

Abou Daya A, Salahuddin MA, Limam N, Boutaba R (2020) Botchase: graph-based bot detection using machine learning. IEEE Trans Netw Serv Manag 17(1):15–29

Aiello LM, Deplano M, Schifanella R, Ruffo G (2012) People are strange when you're a stranger: impact and influence of bots on social networks. In: Proceedings of the international AAAI conference on web and social media, vol 6, pp 10–17

Arin E, Kutlu M (2023) Deep learning based social bot detection on twitter. IEEE Trans Inf Forensics Secur 18:1763–1772

Assenmacher D, Clever L, Frischlich L, Quandt T, Trautmann H, Grimme C (2020) Demystifying social bots: on the intelligence of automated social media actors. Soc Media+ Soc 6(3):2056305120939264

Ayoobi N, Shahriar S, Mukherjee A (2023) The looming threat of fake and llm-generated linkedin profiles: challenges and opportunities for detection and prevention. In: Proceedings of the 34th ACM conference on hypertext and social media, pp 1–10

Cresci S (2020) A decade of social bot detection. Commun ACM 63(10):72–83

Cresci S, Di Pietro R, Petrocchi M, Spognardi A, Tesconi M (2015) Fame for sale: efficient detection of fake twitter followers. Decis Support Syst 80:56–71

Cresci S, Di Pietro R, Petrocchi M, Spognardi A, Tesconi M (2017) 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

Cresci S, Lillo F, Regoli D, Tardelli S, Tesconi M (2018) Fake: evidence of spam and bot activity in stock microblogs on twitter. In: Proceedings of the international AAAI conference on web and social media, vol 12

Cresci S, Lillo F, Regoli D, Tardelli S, Tesconi M (2019a) Cashtag piggybacking: uncovering spam and bot activity in stock microblogs on twitter. ACM Trans Web (TWEB) 13(2):1–27

Cresci S, Petrocchi M, Spognardi A, Tognazzi S (2019b) Better safe than sorry: an adversarial approach to improve social bot detection. In: Proceedings of the 10th ACM conference on web science, pp 47–56

Cresci S, Petrocchi M, Spognardi A, Tognazzi S (2021) The coming age of adversarial social bot detection. First Monday

De Nicola R, Petrocchi M, Pratelli M (2021) On the efficacy of old features for the detection of new bots. Inf Process Manag 58(6):102685

Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805

Dialektakis G, Dimitriadis I, Vakali A (2022) Caleb: a conditional adversarial learning framework to enhance bot detection. arXiv preprint arXiv:2205.15707

Dukić D, Keča D, Stipić D (2020) Are you human? detecting bots on twitter using bert. In: 2020 IEEE 7th international conference on data science and advanced analytics (DSAA). IEEE, pp 631–636

Elyashar A, Fire M, Kagan D, Elovici Y (2016) Guided socialbots: infiltrating the social networks of specific organizations' employees. AI Commun 29(1):87–106

Ester M, Kriegel H-P, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd, vol 96, pp 226–231

Fazil M, Sah AK, Abulaish M (2021) Deepsbd: a deep neural network model with attention mechanism for social bot detection. IEEE Trans Inf Forensics Secur 16:4211–4223

Feng S, Wan H, Wang N, Li J, Luo M (2021) Twibot-20: a comprehensive twitter bot detection benchmark. In: Proceedings of the 30th ACM international conference on information & knowledge management, pp 4485–4494

Feng S, Tan Z, Wan H, Wang N, Chen Z, Zhang B, Zheng Q, Zhang W, Lei Z, Yang S (2022) Twibot-22: towards graph-based twitter bot detection. Adv Neural Inf Process Syst 35:35254–35269

Feng S, Wan H, Wang N, Tan Z, Luo M, Tsvetkov Y (2024) What does the bot say? opportunities and risks of large language models in social media bot detection. arXiv preprint arXiv:2402.00371

Ferrara E (2023) Social bot detection in the age of chatgpt: challenges and opportunities. First Monday

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

Freelon D, Bossetta M, Wells C, Lukito J, Xia Y, Adams K (2022) Black trolls matter: racial and ideological asymmetries in social media disinformation. Soc Sci Comput Rev 40(3):560–578

Freitas C, Benevenuto F, Ghosh S, Veloso A (2015) Reverse engineering socialbot infiltration strategies in twitter. In: Proceedings of the 2015 IEEE/ACM international conference on advances in social networks analysis and mining 2015, pp 25–32

Gilani Z, Farahbakhsh R, Tyson G, Wang L, Crowcroft J (2017) Of bots and humans (on twitter). In: Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017, pp 349–354

Goga O, Venkatadri G, Gummadi KP (2015) The doppelgänger bot attack: exploring identity impersonation in online social networks. In: Proceedings of the 2015 internet measurement conference, pp 141–153

Gorwa R, Guilbeault D (2020) Unpacking the social media bot: a typology to guide research and policy. Policy Internet 12(2):225–248

Grimme B, Pohl J, Winkelmann H, Stampe L, Grimme C (2023) Lost in transformation: rediscovering llm-generated campaigns in social media. In: Multidisciplinary international symposium on disinformation in open online media. Springer, pp 72–87

Guo Q, Xie H, Li Y, Ma W, Zhang C (2021) Social bots detection via fusing bert and graph convolutional networks. Symmetry 14(1):30

Harrag F, Debbah M, Darwish K, Abdelali A (2021) Bert transformer model for detecting Arabic gpt2 auto-generated tweets. arXiv preprint arXiv:2101.09345

Hayawi K, Mathew S, Venugopal N, Masud MM, Ho P-H (2022) Deeprobot: a hybrid deep neural network model for social bot detection based on user profile data. Soc Netw Anal Min 12(1):43

Heidari M, Jones JH (2020) Using bert to extract topic-independent sentiment features for social media bot detection. In: 2020 11th IEEE annual ubiquitous computing, electronics & mobile communication conference (UEMCON). IEEE, pp 0542–0547

Heidari M, Jones JH, Uzuner O (2020) Deep contextualized word embedding for text-based online user profiling to detect social bots on twitter. In: 2020 International conference on data mining workshops (ICDMW). IEEE, pp 480–487

Heidari M, Zad S, Hajibabaee P, Malekzadeh M, HekmatiAthar S, Uzuner O, Jones JH (2021) Bert model for fake news detection based on social bot activities in the covid-19 pandemic. In: 2021 IEEE 12th annual ubiquitous computing, electronics & mobile communication conference (UEMCON). IEEE, pp 0103–0109

Hong L, Luo P, Blanco E, Song X (2024) Outcome-constrained large language models for countering hate speech. arXiv preprint arXiv:2403.17146

Jain D, Arora S, Jha C, Malik G (2024) Text classification models for personality disorders identification. Soc Netw Anal Min 14(1):64

Kennedy I, Wack M, Beers A, Schafer JS, Garcia-Camargo I, Spiro ES, Starbird K (2022) Repeat spreaders and election delegitimization: a comprehensive dataset of misinformation tweets from the 2020 US election. Journal of Quantitative Description: Digital Media 2

Khaliq MA, Chang P, Ma M, Pflugfelder B, Miletić F (2024) Ragar, your falsehood radar: Rag-augmented reasoning for political fact-checking using multimodal large language models. arXiv preprint arXiv:2404.12065

Koonchanok R, Pan Y, Jang H (2024) Public attitudes toward chatgpt on twitter: sentiments, topics, and occupations. Soc Netw Anal Min 14(1):106

Kumarage T, Bhattacharjee A, Garland J (2024) Harnessing artificial intelligence to combat online hate: exploring the challenges and opportunities of large language models in hate speech detection. arXiv preprint arXiv:2403.08035

Lee K, Eoff B, Caverlee J (2011) Seven months with the devils: a long-term study of content polluters on twitter. In: Proceedings of the international AAAI conference on web and social media, vol 5, pp 185–192

Leite JA, Razuvayevskaya O, Bontcheva K, Scarton C (2023) Detecting misinformation with llm-predicted credibility signals and weak supervision. arXiv preprint arXiv:2309.07601

Li S, Zhao C, Li Q, Huang J, Zhao D, Zhu P (2023a) Botfinder: a novel framework for social bots detection in online social networks based on graph embedding and community detection. World Wide Web 26(4):1793–1809

Li S, Yang J, Zhao K (2023b) Are you in a masquerade? exploring the behavior and impact of large language model driven social bots in online social networks. arXiv preprint arXiv:2307.10337

Li G, Lu W, Zhang W, Lian D, Lu K, Mao R, Shu K, Liao H (2024) Re-search for the truth: multi-round retrieval-augmented large language models are strong fake news detectors. arXiv preprint arXiv:2403.09747

Lingam G, Rout RR, Somayajulu DV (2019) Adaptive deep q-learning model for detecting social bots and influential users in online social networks. Appl Intell 49(11):3947–3964

Linvill DL, Warren PL (2020) Troll factories: manufacturing specialized disinformation on twitter. Polit Commun 37(4):447–467

Lopez-Joya S, Diaz-Garcia JA, Ruiz MD, Martin-Bautista MJ (2023) Bot detection in twitter: an overview. In: International conference on flexible query answering systems. Springer, pp 131–144

Lopez-Joya S, Diaz-Garcia JA, Ruiz MD, Martin-Bautista MJ (2024) Exploring social bots: a feature-based approach to improve bot detection in social networks. arXiv preprint arXiv:2411.06626

Mazza M, Cresci S, Avvenuti M, Quattrociocchi W, Tesconi M (2019) Rtbust: exploiting temporal patterns for botnet detection on twitter. In: Proceedings of the 10th ACM conference on web science, pp 183–192

McInnes L, Healy J, Melville J (1802) Umap: uniform manifold approximation and projection for dimension reduction. arxiv 2018. arXiv preprint arXiv:1802.03426

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., Group, P., (2010) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Int J Surg 8(5):336–341

Morstatter F, Wu L, Nazer TH, Carley KM, Liu H (2016) A new approach to bot detection: striking the balance between precision and recall. In: 2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM). IEEE, pp 533–540

Najari S, Salehi M, Farahbakhsh R (2022) Ganbot: a gan-based framework for social bot detection. Soc Netw Anal Min 12(1):4

Nisbet EC, Mortenson C, Li Q (2021) The presumed influence of election misinformation on others reduces our own satisfaction with democracy. The Harvard Kennedy School Misinformation Review

Oentaryo RJ, Murdopo A, Prasetyo PK, Lim E-P (2016) On profiling bots in social media. In: Social informatics: 8th international conference, SocInfo 2016, Bellevue, WA, USA, November 11–14, 2016, proceedings, part I 8. Springer, pp 92–109

Orabi M, Mouheb D, Al Aghbari Z, Kamel I (2020) Detection of bots in social media: a systematic review. Inf Process Manag 57(4):102250

Pearson K (1901) Liii. On lines and planes of closest fit to systems of points in space. Lond Edinb Dubl Philos Mag J Sci 2(11):559–572

Peng H, Zhang J, Huang X, Hao Z, Li A, Yu Z, Yu PS (2024) Unsupervised social bot detection via structural information theory. arXiv preprint arXiv:2404.13595

Pennington J, Socher R, Manning CD (2014) Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543

Pohl JS, Assenmacher D, Seiler MV, Trautmann H, Grimme C (2022) Artificial social media campaign creation for benchmarking and challenging detection approaches. In: ICWSM workshops

Pu J, Sarwar Z, Abdullah SM, Rehman A, Kim Y, Bhattacharya P, Javed M, Viswanath B (2023) Deepfake text detection: limitations and opportunities. In: 2023 IEEE symposium on security and privacy (SP). IEEE, pp 1613–1630

Radivojevic K, Clark N, Brenner P (2024) Llms among us: generative AI participating in digital discourse. In: Proceedings of the AAAI symposium series, vol 3, pp 209–218

Ratkiewicz J, Conover M, Meiss M, Gonçalves B, Flammini A, Menczer F (2011) Detecting and tracking political abuse in social media. In: Proceedings of the international AAAI conference on web and social media, vol 5, pp 297–304

Rauchfleisch A, Kaiser J (2020) The false positive problem of automatic bot detection in social science research. PLoS ONE 15(10):0241045

Richards M, Waugh K, Slaymaker M, Petre M, Woodthorpe J, Gooch D (2024) Bob or bot: exploring chatgpt's answers to university computer science assessment. ACM Trans Comput Educ 24(1):1–32

Rout RR, Lingam G, Somayajulu DV (2020) Detection of malicious social bots using learning automata with url features in twitter network. IEEE Trans Comput Soc Syst 7(4):1004–1018

Sallah A, Arbi Abdellaoui Alaoui E, Agoujil S, Wani MA, Hammad M, Maleh Y, Abd El-Latif AA (2024) Fine-tuned understanding: Enhancing social bot detection with transformer-based classification. IEEE Access 12:118250–118269

Sarzynska-Wawer J, Wawer A, Pawlak A, Szymanowska J, Stefaniak I, Jarkiewicz M, Okruszek L (2021) Detecting formal thought disorder by deep contextualized word representations. Psychiatry Res 304:114135

Sayyadiharikandeh M, Varol O, Yang K-C, Flammini A, Menczer F (2020) Detection of novel social bots by ensembles of specialized classifiers. In: Proceedings of the 29th ACM international conference on information & knowledge management, pp 2725–2732

ShabaniMirzaei T, Chamani H, Abaskohi A, Zadeh ZSH, Bahrak B (2023) A large-scale analysis of Persian tweets regarding covid-19 vaccination. Soc Netw Anal Min 13(1):148

Shi X, Liu J, Song Y (2023) Bert and llm-based multivariate hate speech detection on twitter: comparative analysis and superior performance. In: International artificial intelligence conference. Springer, pp 85–97

Shin D (2023) Algorithms, humans, and interactions: how do algorithms interact with people? Designing meaningful AI experiences. Taylor & Francis, Boca Raton

Shin D (2024) Artificial misinformation: exploring human-algorithm interaction online. Springer, Cham

Shin D, Koerber A, Lim JS (2024) Impact of misinformation from generative AI on user information processing: how people understand misinformation from generative AI. New Media & Society, 14614448241234040

Stieglitz S, Brachten F, Ross B, Jung A-K (2017) Do social bots dream of electric sheep? a categorisation of social media bot accounts. arXiv preprint arXiv:1710.04044

Subrahmanian VS, Azaria A, Durst S, Kagan V, Galstyan A, Lerman K, Zhu L, Ferrara E, Flammini A, Menczer F (2016) The DARPA twitter bot challenge. Computer 49(6):38–46

Varol O, Ferrara E, Davis C, Menczer F, Flammini A (2017) Online human-bot interactions: detection, estimation, and characterization. In: Proceedings of the international AAAI conference on web and social media, vol 11, pp 280–289

Wan H, Feng S, Tan Z, Wang H, Tsvetkov Y, Luo M (2024) Dell: generating reactions and explanations for llm-based misinformation detection. arXiv preprint arXiv:2402.10426

Wang Y, Hamilton AFC (2012) Social top-down response modulation (storm): a model of the control of mimicry in social interaction. Front Hum Neurosci 6:153

Wang H, Hee MS, Awal MR, Choo KTW, Lee RK-W (2023a) Evaluating gpt-3 generated explanations for hateful content moderation. arXiv preprint arXiv:2305.17680

Wang H, Luo X, Wang W, Yan X (2023b) Bot or human? detecting chatgpt imposters with a single question. arXiv preprint arXiv:2305.06424

Wu Y, Fang Y, Shang S, Jin J, Wei L, Wang H (2021) A novel framework for detecting social bots with deep neural networks and active learning. Knowl-Based Syst 211:106525

Wu J, Ye X, Mou C (2023a) Botshape: a novel social bots detection approach via behavioral patterns. arXiv preprint arXiv:2303.10214

Wu J, Ye X, Man Y (2023b) Bottrinet: a unified and efficient embedding for social bots detection via metric learning. In: 2023 11th international symposium on digital forensics and security (ISDFS). IEEE, pp 1–6

Wu Q, Yang Y, He B, Liu H, Wang X, Liao Y, Yang R, Zhou P (2023c) Heterophily-aware social bot detection with supervised contrastive learning. arXiv preprint arXiv:2306.07478

Yang K-C, Menczer F (2023) Anatomy of an AI-powered malicious social botnet. arXiv preprint arXiv:2307.16336

Yang K-C, Varol O, Davis CA, Ferrara E, Flammini A, Menczer F (2019) Arming the public with artificial intelligence to counter social bots. Hum Behav Emerg Technol 1(1):48–61

Yang K-C, Varol O, Hui P-M, Menczer F (2020) Scalable and generalizable social bot detection through data selection. In: Proceedings

of the AAAI conference on artificial intelligence, vol 34, pp 1096–1103

Yang K-C, Ferrara E, Menczer F (2022) Botometer 101: social bot practicum for computational social scientists. J Comput Soc Sci 5(2):1511–1528

Yang Y, Yang R, Peng H, Li Y, Li T, Liao Y, Zhou P (2023) Fedack: federated adversarial contrastive knowledge distillation for cross-lingual and cross-model social bot detection. In: Proceedings of the ACM web conference 2023, pp 1314–1323

Yue Z, Zeng H, Lu Y, Shang L, Zhang Y, Wang D (2024) Evidence-driven retrieval augmented response generation for online misinformation. arXiv preprint arXiv:2403.14952

Zago M, Nespoli P, Papamartzivanos D, Perez MG, Marmol FG, Kambourakis G, Perez GM (2019) Screening out social bots interference: Are there any silver bullets? IEEE Commun Mag 57(8):98–104

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.