Computer Communications 196 (2022) 23–36

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

Investigating the difference between trolls, social bots, and humans on Twitter

Michele Mazza [a,b,*], Marco Avvenuti [a], Stefano Cresci [b], Maurizio Tesconi [b]

[a] Department of Information Engineering, University of Pisa, Via Girolamo Caruso, 16, Pisa, 56122, Italy
[b] Institute of Informatics and Telematics, National Research Council, Via Giuseppe Moruzzi, 1, Pisa, 56127, Italy

ARTICLE INFO

MSC:
0000
1111

Keywords:
Twitter
Information operations
Social bots
Trolls

ABSTRACT

It has become apparent that human accounts are not the sole actors in the social media scenario. The expanding role of social media in the consumption and diffusion of information has been accompanied by attempts to influence public opinion. Researchers reported several instances where social bots, automated accounts designed to impersonate humans, have been deployed for this purpose. More recently, platforms such as Twitter provided evidence pointing to governments creating and using fake accounts in this kind of abuse. These accounts are known as state-backed trolls. Although these different actors have been widely studied, there is little understanding of how they differ when examined together. In this paper, we contribute to understanding the characteristics of the different types of accounts and increase our awareness of Twitter's state-backed trolls, which so far have received limited attention from quantitative researchers. We propose a large-scale quantitative analysis, which relies on both datasets released by Twitter and researchers in recent years to characterize the different actors that take part in the social network scenario. In particular, we represent each account with a large number of features categorized into three distinct traits: credibility, initiative, and adaptability concerning the underlying aspects into which they best fit. We conducted subsequent experiments, isolating features on their respective traits and using them all. First, we apply dimensionality reduction to project accounts onto the same bi-dimensional space and visualize how they distribute across it. Then, we experiment with different combinations of two parameters that affect the dimensionality reduction and clustering algorithm to find which trait is best suited to distinguish the different actors. In our best combination in terms of effectiveness, we obtain high-quality clusters, achieving a purity score of 0.9, which results in homogeneous clusters where accounts of the same category are grouped. Beyond that, we explore our results by visualizing and studying clustering results to determine the differences between account categories. Using our defined traits, we show that it is possible to distinguish the different accounts through clustering, obtaining the best results while leveraging the three traits simultaneously. An additional analysis shows that features related to retweeting patterns and URLs sharing are effective in differentiating trolls and humans. At the same time, social bot accounts suffer from recall degradation in cross-domain evaluation. Moreover, we show that accounts belonging to the same dataset are not necessarily similar in the defined traits. Finally, we perform a feature importance analysis using SHAP to gain insights into which features best differentiate the account when examined in pairs.

1. Introduction

In recent years, social media have become the ground to run campaigns to manipulate public opinion [1]. These campaigns can assume different forms (e.g., disinformation, political propaganda, manipulation campaigns), target individuals and communities, and have various goals [2,3]. To succeed, they require the cooperation of several coordinated accounts to reach a broad audience and obtain a significant impact [4,5]. For this reason, fake accounts are often deployed: social media accounts made of false information or pretending to be a real person or organization [6]. For example, the Internet Research Agency (IRA) created the Twitter account @TEN_GOP named "Tennessee GOP" during its attempt to influence the 2016 U.S. presidential election, attracting more than 100,000 followers [3].

Fake accounts can be fully or partially automated, known as social bots [7] or human-driven, such as trolls. Previous work has focused mainly on the activity of social bots [8]; however, automated accounts are only part of the problem related to content diffusion. Since the

[*] Corresponding author at: Department of Information Engineering, University of Pisa, Via Girolamo Caruso, 16, Pisa, 56122, Italy.
E-mail address: michele.mazza@iit.cnr.it (M. Mazza).

https://doi.org/10.1016/j.comcom.2022.09.022
Received 8 December 2021; Received in revised form 25 July 2022; Accepted 22 September 2022
Available online 27 September 2022

release of the first dataset containing accounts attributed to state-linked information operations (IOs), state-backed trolls have become the focus of extensive scientific attention. These accounts are created and used by governments to carry out propaganda and disinformation campaigns. However, they are not the only kind of human actor that may be involved in these malicious campaigns, as also unaware humans can play a crucial role in their outcome [8]. Researchers and social media platforms have recently been active in seeking solutions to the many problems that affect social media [9,10], including those caused by fake accounts by developing methods and techniques to identify the fakes and their malicious campaigns [7,11]. However, fake accounts campaigns are still considered a threat to societies and democracies [12]. The most recent efforts to study campaigns conducted by fake accounts focused on shared content, for example, narratives, their veracity, and their spreading patterns in and across platforms. Moreover, investigations of new and large-scale malicious campaigns require a great deal of manual, qualitative, or mixed-method analyses to verify claims, trace the sources of mis/disinformation, and perform attribution. This implies that limited attention has been devoted to comprehensive and extensive large-scale quantitative analyses of the different accounts that may be involved. Compared to the vast literature regarding malicious and deceptive social bot accounts, state-backed trolls have received limited scholarly attention [13]. Overall, many aspects of the different actors remain unclear, e.g., from a computational and quantitative standpoint, what are the differences (if any) between the online behaviors of social bots, humans, and trolls? On the basis of the features identified in the literature to analyze social network accounts, is it possible to identify groups of features that can help us distinguish these three types of accounts? Are the characteristics shown by the accounts consistent across datasets released in the past years? We aim to address these questions by relying on Twitter's IO datasets and the most comprehensive bot repository that exists to date. We leverage such extensive data by adopting state-of-the-art techniques from social media analysis, natural language processing and unsupervised machine learning.

1.1. Contributions

We contribute to filling the scientific gap regarding the differences between human accounts, social bots, and state-backed trolls through a large-scale quantitative analysis. For our analysis, we compute 99 features out of public Twitter data about both the accounts and the content they produce. We associate each feature to a particular trait, defined as "a distinguishing quality or characteristic", according to which we analyze the different accounts. In particular, we leverage recent theoretical and empirical findings [1,13–17] and highlight three traits of accounts: *credibility*, *initiative*, and *adaptability*. From the recent literature, we expect the different types of accounts to behave differently across these three traits [1,7]. We verify this assumption by applying unsupervised dimensionality reduction and density-based clustering to our fine-grained features. We analyze our results qualitatively by visualizing and studying how the accounts are positioned in the bidimensional space according to the traits identified and quantitatively — by leveraging well-known metrics such as cluster purity. As has already occurred with social bots [14,15,15,18–20], a segment of research has begun to develop automatic techniques in an effort to detect troll accounts [16,21,22]. These works analyze only the accounts belonging to the IRA dataset released by Twitter, while here, we analyze troll accounts from several datasets. Moreover, existing works regarding fake account detection focus only on the binary distinction between trolls and humans or between social bots and humans. Although detection algorithms might be generalized to a broader domain, a systematically detailed analysis of the differences between the three types of accounts has not yet been done. Understanding how trolls differ from social bots and humans can help those who will work on their identification to better characterize them. Our main contributions are as follows:

- We built a massive data set using Twitter IO data sets and the Botometer Bot Repository, which consists of a collection of annotated datasets of social bots and humans. To the best of our knowledge, no work has ever been done on such kind of dataset.
- We represented the accounts through 99 features concerning both the profile and the content they produce and bound them to three traits accounting for different extents an account can behave.
- We performed experiments on both the single trait and their combination, demonstrating that together they are more effective for identifying groups of accounts of the same type, with cluster purity exceeding 0.9 on a 0–1 scale.
- We showed that dimensionality reduction techniques, such as UMAP, followed by HDBSCAN clustering, can be used effectively to group accounts of the same type.
- We deepened our results by analyzing the overall distribution of the different actors across clusters. This is achieved by studying differences in terms of features among the two largest groups of humans and trolls and finding resonances with the recall decay phenomenon in cross-domain social bot detection.
- We performed a feature importance analysis using SHAP to gain insights into which features better differentiate between different types of accounts.

2. Related work

Most literature examining fake accounts has focused on social bot detection. There is evidence that social bots are involved in campaigns to influence public opinion [7,23] as well as to spread fake news [7] and conspiracy theories [24]. Recently, social bots have manipulated the stock market and infiltrated political discourse [25,26]. The development of methods for their detection has revealed several aspects of their behavior. Social bots can be distinguished from human accounts according to their profile [15,20], their activity [15,18], the shared content [14,27], and their social or interaction graph [14,19].

Concerning the representation of accounts by features, [28] proposed a framework that evaluates the characteristics of a Twitter account to determine if it is a social bot. The framework leverages an ensemble classifier based on feature engineering that aims to provide an indicator, namely, a bot score, to classify an account as a bot. Choosing the relevant features to describe the entities to be classified is a crucial step in machine learning classifiers. The literature considered a wide variety of choices; however, in general, six main categories of features were proposed to identify social bot accounts [14]: user metadata, friend metadata, retweet/mention network structure, content and language, sentiment, and temporal features. As for less advanced social bots, a user can infer the profile inauthenticity by checking suspicious profile information, such as automatically-generated usernames, unbalanced following-follower ratios, recent profile creation times, and missing or inconsistent profile descriptions [28,29]. In other cases, social bots can be distinguished more easily because they are designed for a single purpose, such as increasing the number of followers or retweets of an account [30,31]. For example, the technique discussed in [32] builds a representation inspired by DNA sequencing and based on the sequence of interactions of each account, which has been shown to be adequate for distinguishing between human accounts and spambots. Recent studies have shown that social bots are becoming more sophisticated: new generations of social bots mimic humans better than ever before [7,33]. For this reason, social bot detection and characterization are becoming increasingly complex tasks. In contrast, state-backed trolls are set up without software automation and operated manually. Currently, the consensus is that the close similarity between the troll and human accounts makes it much more difficult to recognize them through automated detection techniques [16,22,23,34], which are better suited to identify automated behaviors. As far as we know, a limited number of works studied trolls features to build automatic detection techniques [16,21,22] and all of them are built only on the

IRA dataset released by Twitter. However, data on accounts involved in IOs released by Twitter have triggered a round of quantitative research, especially on Russian trolls, suggesting that troll behavior differs significantly from human accounts in terms of posting patterns and language use [13,16,35]. Exploratory studies characterized trolls along with different behaviors such as hashtags, URLs, and retweet patterns [13,36]. In particular, trolls often show anomalous tweet and retweet rates while sharing more URLs and using more hashtags [16, 35]. Additionally, compared to those shared by human accounts, troll tweets appear shorter and consist of shorter words [37]. Trolls have also been classified according to their actions: [38] identified three groups of trolls based on their behavior: right trolls, left trolls, and news feed trolls, while [39] introduced two other groups: hashtag gamer and fearmonger. Finally, [40] pointed out that their strategic behavior changes over time regardless of which group they belong to. Another research branch tries to identify fake accounts by studying their correlation with malicious content on social networks. Typically, this is done using blacklists provided by professional fact-checkers who manually assess the nature of content shared by accounts. Despite being widely adopted, this approach can scarcely keep up with the dynamic strategies of fake accounts [34]. Unlike previous work, our study does not focus solely on one type of account, and our characterization is extended to three categories of accounts. To the best of our knowledge, this constitutes the first effort of a large-scale quantitative analysis comprising both datasets officially released by Twitter and datasets acquired by researchers studying social bots over the last years.

3. Datasets

To model and characterize the different accounts on Twitter, we first constructed a large dataset of trolls, social bots, and human accounts, as summarized in Table 1. We obtained data on the social bot and human accounts through datasets made available by researchers in recent years and collected in the Bot Repository.[1] Since these datasets only report the ID and the label assigned to the relative user, we exploited the Twitter API to rehydrate user activity and profile information. However, we could not obtain the data for the suspended and deleted accounts, so the numbers shown in Table 1 might be smaller than those of the original papers. Furthermore, according to the Twitter API limits, we collected a maximum of 3.200 tweets per account. Most of the datasets available on the Bot Repository come from efforts to identify or characterize social bots. Although many accounts have been manually annotated by humans, some have been created through meta-data filtering or automated techniques. The `verified-2019` dataset was obtained by filtering the streaming API for verified accounts [20]. The `botwiki-2019` was created exploiting the botwiki.org archive, a catalog of self-identified social bots accounts [20]. Social bots and humans in `cresci-rtbust-2019` were manually annotated by the authors from a much larger dataset [18]. The `political-bots` consist of social bots that show political orientation, shared by the user `@josh_emerson` [31]. Accounts in `botometer-feedback-2019` were first flagged by Botometer users and then manually annotated [31]. The `vendor-purchased-2019` is made of fake followers purchased by the CNetS team, who also collected celebrities accounts contained in `celebrity-2019` [31]. The `pronbots-2019` was first shared by Andy Patel[2] and consists of social bots sharing scam websites [31]. In `midterm-2018`, human accounts were manually identified, while social bots were spotted through suspicious correlations in their creation and tweeting timestamps [20]. Social bots in the `cresci-stock-2018` dataset were detected by finding accounts with similar timelines [26]. In the `gilani-2017` dataset, accounts were annotated following key information compiled in a table [41]. For the dataset `varol-2017`, accounts sampled from different Botometer score deciles were manually annotated [14]. Inside the `cresci-2017` dataset are identified three different classes of social bot: traditional spambots, social spambots, and fake followers [33]. These social bots, designed for different purposes, vary in their behavior: traditional spambots share tons of content, social spambots tend to interact mainly with political candidates, and fake followers have aggressive following patterns [28]. We have unpacked the dataset into smaller datasets to preserve this distinction and better evaluate our results. The `cresci-2015` dataset has also been unpacked since it contains human accounts who voluntarily joined “The Fake Project”, placed in the dataset `thefakeproject-2015`, and human accounts involved in politics [42] which are in the dataset `elections-2013`.

Regarding troll accounts, we leaned on the official Twitter datasets.[3] The datasets have been released to allow a further investigation by those who do not work directly with the company. This resulted in a repository of datasets concerning potential foreign IOs that occurred on Twitter, containing the complete timeline of these state-backed trolls, including shared media. Although Twitter has provided each user’s activity since its creation, we have cut up to the 3.200 most recent tweets for each account to balance the data available on trolls with that for humans and social bots we obtained by rehydrating users’ activity through Twitter APIs. We cannot summarize how these datasets have been collected, as it is not publicly how Twitter identifies IOs and the involved accounts. The dataset names indicate where the IO originated. The accounts contained in multiple datasets have been counted only once, resulting in a dataset containing 91,632 distinct accounts:

- 27,877 humans;
- 32,742 social bots;
- 31,013 trolls.

As far as we know, this is the first work analyzing a massive dataset obtained by merging both datasets released by research and Twitter, containing these three types of social network accounts.

4. Method

4.1. Features binding

Our study leverages data related to the user profile and tweets/ retweets shared by the user obtained from the Twitter REST API[4] for social bot and human accounts. For state-backed trolls, the data are directly provided by Twitter. We distilled the data and meta-data into 99 features, listed in Table 2, each of which is represented as a continuous numeric value, a binary value, or a set of statistics computed from distribution and used as individual features: minimum, maximum, median, mean, standard deviation, skewness, and entropy. We selected features that have been proven to be effective in distinguishing social bots and human accounts [15,42], many of which have also been used to identify trolls [16]. In addition, we introduced a new feature regarding language novelty, which measures the percentage of new tokens in a tweet compared to those used in previous tweets. This metric measures how much an account’s lexicon grows over time in a language-independent manner and is deemed helpful due to recent evidence on the time-varying behavior of state-backed trolls [22,36, 40].

Typically, account features are loosely grouped into broad categories based on their domains. They have been arranged into categories such as user-based, friends, network, temporal, content, and sentiment [14,15]. Other works focused on a particular domain and increased its granularity. [16] split the language domain into stop-word usage, language distribution, and bag-of-words features. Even features extracted in this work could be grouped into similar domains. However, we pursued a more intuitive grouping based on the different extents an

[1] https://botometer.iuni.iu.edu/bot-repository/

[2] github.com/r0zetta/pronbot2

[3] https://transparency.twitter.com/en/reports/information-operations.html

[4] https://developer.twitter.com/en/docs

Table 1
List of annotated datasets used for experiments.

Dataset	Human	Social Bot	Troll	Total
Bot Repository				
verified-2019 [20]	1,980	–	–	1,980
botwiki-2019 [20]	–	662	–	662
cresci-rtbust-2019 [18]	329	317	–	646
political-bots-2019 [31]	–	13	–	13
botometer-feedback-2019 [31]	341	111	–	452
vendor-purchased-2019 [31]	–	747	–	747
celebrity-2019 [31]	5,821	–	–	5,821
pronbots-2019 [31]	–	1,738	–	1,738
midterm-2018 [20]	7,628	45	–	7,673
cresci-stock-2018 [26]	7,474	18,508	–	25,982
gilani-2017 [41]	1,350	994	–	2,344
varol-2017 [14]	1,008	493	–	1,501
cresci-2017 - social_spambots_1-2017 [33]	–	991	–	991
cresci-2017 - social_spambots_2-2017 [33]	–	3,457	–	3,457
cresci-2017 - social_spambots_3-2017 [33]	–	464	–	464
cresci-2017 - traditional_spambots-2017 [33]	–	1,000	–	1,000
cresci-2017 - fake_followers-2017 [33]	–	3,202	–	3,202
cresci-2015 - thefakeproject-2015 [42]	465	–	–	465
cresci-2015 - elections-2013 [42]	1,481	–	–	1,481
Twitter				
Saudi Arabia (October 2019)	–	–	5,929	5,929
China (July 2019, set3)	–	–	4,301	4,301
Saudi Arabia (April 2019)	–	–	6	6
Ecuador (April 2019)	–	–	1,019	1,019
UAE (March 2019)	–	–	4,248	4,248
Spain (April 2019)	–	–	259	259
UAE/Egypt (April 2019)	–	–	271	271
China (July 2019, set 1)	–	–	744	744
China (July 2019, set 2)	–	–	196	196
Catalonia (June 2019)	–	–	130	130
Iran (June 2019, set 1)	–	–	1,666	1,666
Iran (June 2019, set 2)	–	–	248	248
Iran (June 2019, set 3)	–	–	2,865	2,865
Russia (June 2019)	–	–	4	4
Venezuela (June 2019)	–	–	33	33
Iran (January 2019)	–	–	2,320	2,320
Bangladesh (January 2019)	–	–	15	15
Russia (January 2019)	–	–	416	416
Venezuela (January 2019, set 1)	–	–	1,196	1,196
Venezuela (January 2019, set 2)	–	–	764	764
IRA (October 2018)	–	–	3,613	3.613
Iran (October 2018)	–	–	770	770
Distinct users	**27,877**	**32,742**	**31,013**	**91,632**

account can behave in the social network scenario. We followed and reworked the theoretical and empirical findings, highlighting attributes characterizing the different types of accounts. These relate to the trustworthiness of the account [1], the topics it discusses [13], the way its behavior evolves in different contexts [13,17] and the strategies it adopts to achieve its goals [13]. This investigation led to the definition of three groups of features, named by us as traits, appropriate to represent aspects through which research today describes and distinguishes the different social network accounts:

- **credibility**: measures the extent to which an account is credible and trustworthy, based on known characteristics of low-credibility accounts (e.g., number and type of social relationships, account age, productivity, etc.);
- **initiative**: concerns the degree to which an account is capable of sparking new debates, driving the conversation, and contributing original content, or whether it just follows and reshares what others said (e.g., the ratio between original and reshared content, the ratio between tweets and replies, etc.);
- **adaptability**: is related to how an account modifies and adapts its profile and behavior to different times, and topics it is exposed/contributes to (e.g., language novelty, entropy, and diversity in the account behavior, etc.).

We attributed to the credibility trait the features strictly related to a user profile, i.e., those that even an ordinary user can assess while observing the account directly on a social network. The initiative includes features around the type and quantity of shared content, while adaptability is related to the time-varying characteristics of the account and its linguistic features. Our three traits can be traced back to various approaches provided by the research community to detect or characterize social bots and state-backed trolls. [42] leveraged credibility-based features to identify so-called fake followers. In [20], features that we associate with credibility, such as the number of followers and favorites, were found to be sufficient to distinguish human and social bot accounts. [16] labeled three categories to identify active state-backed trolls: profile features, behavioral features, and language distribution features. These categories can be considered subgroups of our three traits: profile features such as account age or the number of followers are found in credibility, behavioral features related to shared URLs are attributed to the initiative, and language distribution features belong to adaptability.

4.2. Dimensionality reduction

Since our goal is to find groups of similar accounts through clustering, we must address the problem regarding the number of features. As the dimension increases, the distance from any point to the nearest data

Table 2
List of features extracted for each trait.

Feature	Type	Description
Credibility		
Friends count	Numeric	Number of friends
Followers count	Numeric	Number of followers
Favorites ratio	Numeric	Ratio between favorites received and tweets
Followers ratio	Numeric	Ratio between the number of friends and the number of followers squared
Has URL	Binary	If profile reports a URL
Bio sentences	Numeric	Number of sentences in bio
Bio tokens	Numeric	Number of tokens in bio
Bio characters	Numeric	Number of characters in bio
Has bio	Binary	If profile reports a bio
Account age	Numeric	Account age expressed in days
Following ratio	Numeric	Ratio between friends and age
Tweet ratio	Numeric	Ratio between tweets and age
Languages	Numeric	Number of distinct languages found in tweets
Languages Ratio	Numeric	Ratio between the number of languages and age
Sources	Numeric	Number of distinct sources
Sources ratio	Numeric	Ratio between the number of sources and age
API ratio	Numeric	Ratio between the number of custom and standard sources
Initiative		
Retweet ratio	Numeric	Ratio between the number of retweets and the number of tweets
Reply ratio	Numeric	Ratio between the number of replies and the number of tweets
Tweet-URL ratio	Numeric	Ratio between the number of tweets containing a URL and the number of tweets
Retweet-URL ratio	Numeric	Ratio between the number of retweets containing a URL and the number of retweets
Reply-URL ratio	Numeric	Ratio between the number of replies containing a URL and the number of retweets
Words in tweets	Distribution parameters	Distribution of the number of unique words in tweets
Words entropy in tweets	Distribution parameters	Distribution of the number of unique words entropy in tweets
Adaptability		
Languages	Distribution parameters	Distribution of the number of languages used
Language Novelty	Distribution parameters	Percentage of new tokens in a tweet compared to those used in the previous
Time between tweets	Distribution parameters	Distribution of time differences between consecutive tweets
Time between retweets	Distribution parameters	Distribution of time differences between consecutive retweets
Time between mentions	Distribution parameters	Distribution of time differences between consecutive tweets containing mentions
Retweeted accounts	Distribution parameters	Distribution of the number of retweeted accounts
URL domains	Distribution parameters	Distribution of the number of domains contained in tweets
Tweets words	Distribution parameters	Distribution of the number of words contained in tweets
Tweets characters	Distribution parameters	Distribution of the number of characters contained in tweets

point has been shown to be closer to the distance to the closest data point [43]. This is problematic for clustering techniques, which typically assume short within-cluster and large between-cluster distances. To address this issue, we relied on uniform manifold approximation and projection (UMAP) [44] to reduce the cardinality of our vectors of features. UMAP, at its core, operates very similarly to t-SNE [45] since both use graph layout algorithms to organize data in a low-dimensional space. The concept behind UMAP is quite simple: it builds a high-dimensional graph representation of the data and then optimizes a low-dimensional graph to be as structurally similar as possible. This dimensionality reduction technique has become popular due to its tendency to reveal clusters. Moreover, it is significantly more computationally efficient than other well-known dimensionality reduction algorithms such as t-SNE [44].

4.3. Clustering

Having each account represented by a feature vector within the low-dimensional space projected by UMAP, we applied density-based clustering to check the effectiveness of the traits in generating clusters of accounts of the same type. We base our clustering step on an efficient algorithm that combines density- and hierarchical-based clustering: HDBSCAN [46]. HDBSCAN extends DBSCAN [47] by converting it into a hierarchical clustering algorithm and then using a technique to extract a flat clustering based on the stability of the clusters. HDBSCAN basically iterates DBSCAN over several values of the parameter ϵ, which defines how close points should be to each other to be considered part of the same cluster, and integrates the results to find a clustering that provides the best stability over ϵ. This allows HDBSCAN to find clusters of varying densities and to be more robust to parameter selection. Among the advantages of this algorithm is its effectiveness in finding clusters with variable degrees of density, a much-desired feature when dealing with noisy real-world data. In addition, HDBSCAN also proved about twice as fast as its predecessor, DBSCAN [47]. Regarding the algorithm parameters, HDBSCAN does not require specifying a global density threshold, the parameter ϵ in DBSCAN, by employing the optimization strategy described before.

4.4. Experiments

The algorithms we used for dimensionality reduction and clustering allow the regulation of parameters that can significantly affect the final result. We considered the most impactful parameters for both algorithms on the outcome, which result in $num\_components$ for UMAP and $min\_clusters\_size$ for HDBSCAN. The parameter $num\_components$ affects the dimensionality of the reduced dimension space in which we embed the data, while $min\_clusters\_size$ affects the smallest size grouping we consider as a cluster. Since we could not determine the best combination, we experimented by varying and combining different values for these parameters. Through the parameter $num\_components$, we operate on the data sparsity since, by reducing the size of the feature vectors, the data result is essentially unevenly spaced. The $min\_cluster\_size$ parameter is exploited to reduce the number of resulting clusters, as the HDBSCAN algorithm tends to merge different clusters when the parameter increases. Considering that each trait has a different number of features, we selected $num\_components$ by dividing the interval between 2 and the total number of features into nine bins. For $min\_cluster\_size$, we set a maximum value of 893, corresponding to 1% of the entire dataset. The remaining values were obtained by dividing each time by two until we obtained three as the minimum value. We tested each *min_cluster_size* value combined with different *num_components* values for each of the three traits, for a total of 243 experiments.

4.5. Purity

To estimate the effectiveness of our technique, we leveraged our prior knowledge of the dataset with external validity criteria. This evaluation compares the clustering results with the labeled ground truth of the dataset [48]. We chose the purity score over other measures such as Rand Index (RI) and Normalized Mutual Information (NMI) because it does not penalize accounts of different categories in the same cluster (false positives) but measures the accuracy of the cluster assignment by counting the number of correctly assigned elements and dividing by the cluster cardinality, keeping the evaluation measure simple and transparent. In addition, to compute purity, each cluster is assigned to the most frequent class in the cluster, avoiding the need to manually assign a class to each cluster. The only drawback of the purity score is that it is easy to achieve a high score when the number of clusters is significant. However, we overcome this problem by assigning incremental values to the parameter `min_cluster_size` for the HDBSCAN algorithm. The purity score value ranges from 0 to 1, where a high value means that elements of the class tend to cluster together. Therefore, we calculated the mean purity score of the set of clusters obtained from each experiment. Formally, given N the number of objects, k the number of clusters, c_i a cluster in C, and t_j the classification that has the maximum count for the cluster c_i, the purity score is defined as:

$$Purity = \frac{1}{N}\sum_{i=1}^{k} max_j |c_i \cap t_j|$$

Purity does not work well for unbalanced data since even poorly performing clustering algorithms will give a high purity value [49]. However, since our dataset is relatively well balanced, the purity scores obtained in our experiments are reliable. For our experiments, we considered all unclustered points[5] to belong to the same cluster.

4.6. Silhouette

We used an additional metric, the silhouette coefficient, to identify the best clustering set among those obtained. The silhouette coefficient, a metric used to interpret and validate consistency within clusters of data, measures how similar an object is to its cluster (cohesion) compared to other clusters (separation). Its value ranges from −1 to +1, where a high value indicates that the clusters are well apart from each other and clearly distinguished; values near 0 mean that the clusters are indifferent, or more precisely, that the distance between the clusters is not significant. In contrast, a low value corresponds to clusters that are assigned incorrectly. To calculate the silhouette coefficient, one needs to calculate, for each point $i \in C_i$, the distance between i and all other points in its cluster. Considering $|C_i|$ the cardinality of the cluster:

$$a(i) = \frac{1}{|C_i| - 1}\sum_{j \in C_i, i \neq j} d(i,j)$$

We define the mean dissimilarity of point i to some cluster C_k as the mean of the distance from i to all points in C_k (where $C_k \neq C_i$). For each data point $i \in C_i$, we calculate:

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|}\sum_{j \in C_k} d(i,j)$$

By leveraging the previous metrics we can define the silhouette index for a given point i as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad if \quad |C_i| > 1$$

The maximum value of the mean $s(i)$ over all data of the entire dataset is the silhouette coefficient:

$$Silhouette \ \ Coefficient = \max_k \tilde{s}(k)$$

Where $\tilde{s}(k)$ represents the mean $s(i)$ of all data in the entire dataset for a specific number of clusters k. We exploited the silhouette coefficient to choose the cluster set to explore among the results obtained by combining the three traits.

4.7. Jaccard Similarity Index

We used the Jaccard similarity index to calculate the similarity between two clusters resulting from different experiments. Measuring the Jaccard Similarity Index between two clusters is the result of the division between the number of elements they have in common, as shown below:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Two clusters that share all elements will have Jaccard Similarity Index 1, the closer to 1, the more elements are in common between the clusters, while if they share no elements, the Jaccard Similarity Index would be 0.

5. Results

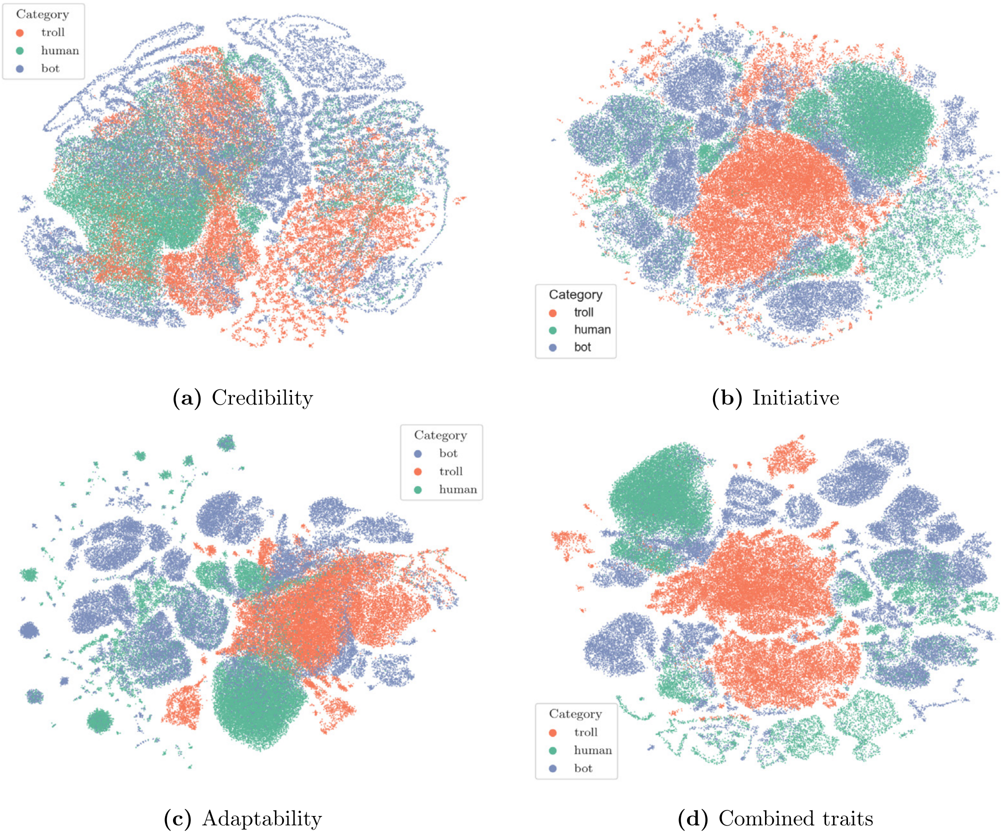
5.1. Qualitative analysis

First, we performed a qualitative analysis applying UMAP with *num_components* = 2 on each set of features to gain an initial insight into how accounts are distributed in bidimensional space. In Fig. 1, the accounts are colored according to their nature (e.g., bot, troll, or human) and plotted for each trait and their combination. Using colors and the distribution of points, it is possible to visually assess the effectiveness of the features extracted for each trait in distinguishing different actors by observing the proximity of points representing accounts of the same type. Concerning credibility (Fig. 1(a)), we observe that the different categories tend to overlap, except in a few rare areas. Regarding initiative (Fig. 1(b)), groups of accounts of the same type are visible, although they are all very close to each other, making the distinction into clusters difficult. Adaptability (Fig. 1(c)), while providing well-separated groups of accounts of the same category, also creates a vast group where most human and troll accounts are located. Combining the three traits (Fig. 1(d)) will result in more groups of accounts belonging to the same category that are well separated from each other. Here emerge two groups of troll accounts that stand out from the rest, a large group of human accounts and other small well-delineated groups.

5.2. Quantitative analysis

After preliminary data exploration, we performed a battery of experiments using different values for the parameters *num_components* and *min_clusters_size*, respectively, for the UMAP and HDBSCAN algorithms. In each experiment, we analyzed all 91,632 accounts in the dataset according to the group of features bundled into the different traits. Since our dataset contains three different types of accounts, the number of generated clusters would ideally be three. However, since the accounts analyzed belong to various datasets and performed different activities at different times, they probably do not exhibit the same characteristics even though they belong to the same category. We report the distribution of the purity values obtained in our experiments in Fig. 2 through box plots that are consistent with what was observed in Fig. 1. Credibility gives the worst results; its purity values fall below 0.5 and do not exceed 0.8, while the purity values for initiative and adaptability are comparable, although the values for initiative always remain above 0.7. The best results of the purity score are obtained by combining the features of all identified traits.

[5] In density-based clustering, points laying in a region with low point density are not added to any cluster.



(a) Credibility

(b) Initiative

(c) Adaptability

(d) Combined traits

Fig. 1. Visual exploration of the features extracted in a bidimensional space in the different traits and their combination. Each point represents an account, colored according to its nature (bot, troll, human). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

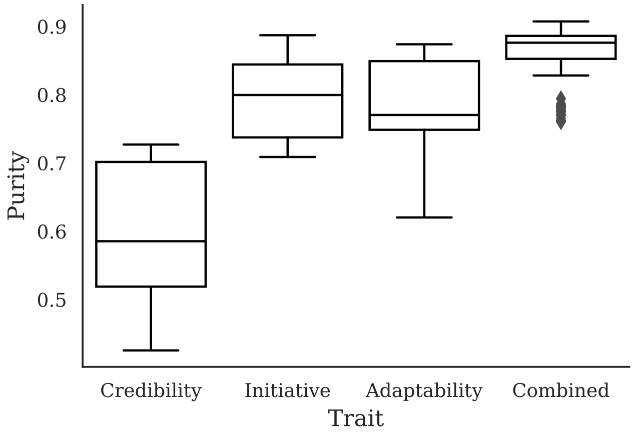


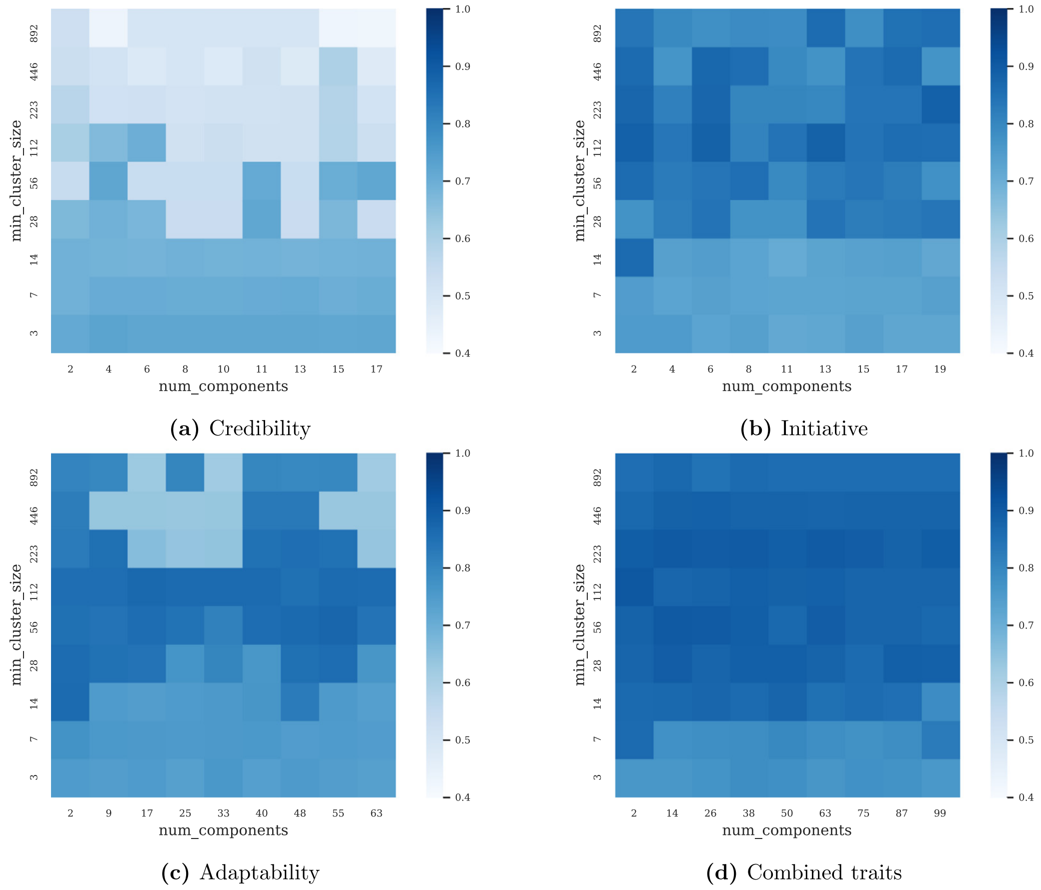
Fig. 2. Box plot relating to the purity scores obtained for each trait.

Fig. 3 shows in detail the purity score resulting from each experiment. Credibility, Fig. 3(a), is the trait with lower purity values, as can be inferred from Fig. 1(a) not showing a proper separation between different categories of accounts. Previous research highlighted the effectiveness of social bots [42] and state-backed trolls [3] in impersonating and imitating human accounts. This explains why the features that belong to credibility are insufficient to distinguish different types of accounts. The initiative, Fig. 3(b), generally returns higher purity values than other traits, which do not fall below 0.7 and increase as the parameter *min_cluster_size* increases. Features belonging to the initiative are those most related to the activity of an account and hence more difficult to disguise and blend with those of a human account. Moreover, since state-backed troll's IOs show more complex dynamics than the campaigns in which bots are exploited, this trait effectively distinguishes them. The adaptability trait, Fig. 3(c), is effective for some values of *min_cluster_size*, in particular 56 and 112, with little variation from *num_components*. The effectiveness of this trait depends significantly on the selected parameters, while its purity score is not linear, as in the other traits. Figs. 2 and 3(d) show that using all traits, we obtain a purity value greater than 0.8 in most setups. This indicates that although the different accounts behave differently in different contexts, they show similarities in at least one identified trait.

Furthermore, the results of the experiments suggest that the parameter *min_cluster_size* significantly affects the purity score more than the parameter *num_components*. As *min_cluster_size* increases, the purity score increases, except for credibility, where purity decreases as *min_cluster_size* increases. As we increase *min_cluster_size*, we obtain fewer clusters; this trend contrasts with the notion of purity scores being higher in the case of several clusters. Clusters aggregated when *min_cluster_size* is increased are likely to be mainly composed of the same account category.

5.3. Accounts distribution

We chose one configuration of parameters among several used in the experiments with the combined traits to understand how the various accounts are distributed across the clusters. In Fig. 4, we plotted every cluster set against its silhouette coefficient on the y axis and its purity against the x axis. Among the various cluster sets, we chose the one with a higher purity score and silhouette coefficient. This set corresponds to the experiment carried out with $min\_cluster\_size = 223$ and $num\_components = 63$, highlighted in Fig. 4, which returns 38 clusters with a purity of 0.9 and a silhouette coefficient of 0.77. Before further investigation, we observed the distribution of the clusters in the cluster sets surrounding the one we selected, which also show high purity and silhouette values, shown in Fig. A.11.



(a) Credibility

(b) Initiative

(c) Adaptability

(d) Combined traits

Fig. 3. Heatmaps reporting the purity values obtained for each set of experiments for the different traits and their combination. Darker shades correspond to higher purity values. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

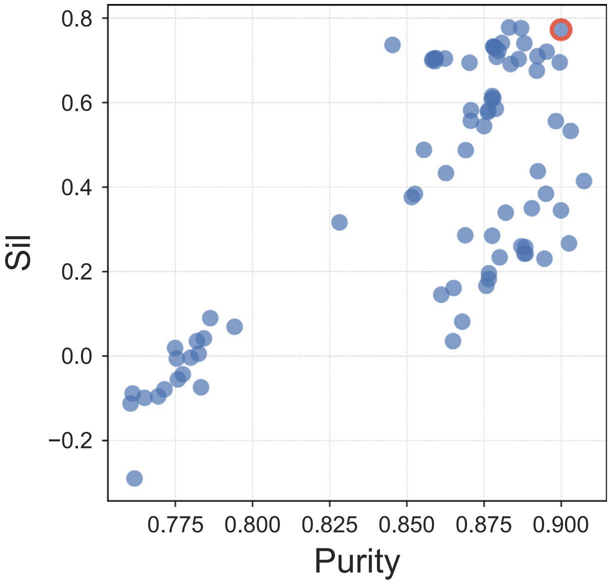


Fig. 4. Points are plotted against the silhouette coefficient on the y axis and the purity score on the x axis. Each point represents a set of clusters obtained from different experiments performed using the combined three traits. The best experimental condition is highlighted in the figure.

The largest clusters, particularly the first two, tend to always form in the same manner: one is mainly composed of trolls, while humans mostly form the other. Only two experiments produce slightly different results (marked in Table 3 and Fig. A.11), consisting of a group of social bots aggregated to the cluster containing most humans. As for social bots, they are more sparse among the different clusters. Then, we explored how the three account categories were distributed among obtained clusters in the chosen cluster set. Fig. 5 shows that about 55% of human accounts are grouped in cluster 2. Regarding troll accounts, most of them are allocated between two clusters. In cluster 1, we find about 56% of the analyzed troll accounts, while in cluster 3 are located slightly less than 27%. Furthermore, our results suggest that for accounts collected in the literature and those released by Twitter, state-backed trolls are easily distinguishable from the social bot and human accounts. This outcome is in contrast to the consensus that trolls are hard to discern from humans since both are controlled by a human agent [16,22,23,34]. There are several possible explanations for this. An explanation concerns the underlying background of the data used in our study: the collected datasets come from various sources and contain accounts that operated in diversified environments. In particular, human and social bot accounts come from datasets collected by research, mainly in the same work, while troll accounts are directly provided by Twitter. This cross-domain setting may have facilitated the distinction of troll accounts. Another explanation could be that all previous work focuses primarily on the Twitter IRA dataset. In contrast, in our work, it constitutes less than 12% of the troll accounts examined since we included many other datasets. However, our results provide a starting point for a better understanding of the currently available datasets that represent the accounts that operate on social networks.

To ensure that the results obtained are not biased, for example, by the presence of prominent features, such as the account age arising from the time sensitiveness of the datasets, we furthered our analysis on clusters 1 and 2. As we intended to study the differences between the

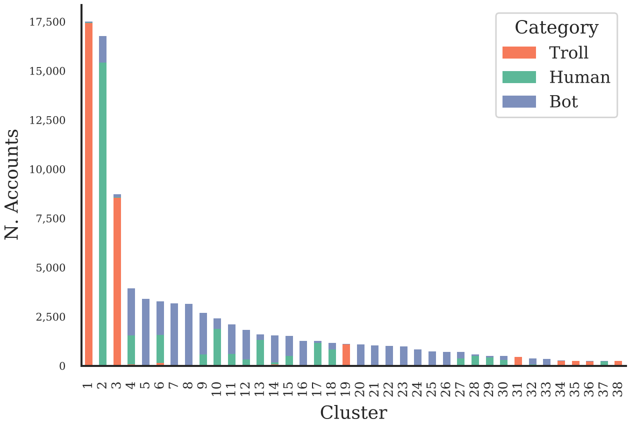


Fig. 5. Distribution of the three different categories of accounts over the 38 clusters obtained in the experimental setup with $min\_cluster\_size = 223$ and $num\_components = 63$.

Table 3
Jaccard similarity between the two largest clusters of the selected cluster set and the remaining 17 candidates.

min_cluster_size, num_components	Cluster 1	Cluster 2
∗112, 50	0.99	0.74
223, 75	0.99	0.99
223, 87	0.99	0.99
223, 99	0.99	0.99
446, 26	0.99	0.99
446, 38	0.99	0.99
446, 50	0.99	0.99
446, 63	0.99	0.99
446, 75	0.99	0.99
446, 87	0.99	0.99
446, 99	0.99	0.99
∗892, 26	0.97	0.72
892, 38	0.99	0.98
892, 63	0.99	0.98
892, 75	0.99	0.98
892, 87	0.99	0.98
892, 99	0.99	0.98

two clusters, in Table 3, we computed the Jaccard similarity between the two most significant clusters of the selected cluster set and the surrounding ones shown in Fig. 4 that also exhibit both high purity and silhouette coefficient. The high Jaccard similarity values obtained assure us that the chosen set is representative.

5.4. Differences between humans and trolls

We investigated which features differentiate the accounts belonging to clusters 1 and 2 of the selected experiment, taking advantage of their similar cardinality, meaning that they have a comparable number of elements, respectively 17,482 and 16,771, each dominated by different types of accounts, trolls in cluster 1 and humans in cluster 2.

First, we filtered out non-troll accounts from cluster 1 and nonhuman accounts from cluster 2, obtaining two new clusters: cluster $1a$ composed of 17,439 trolls account and cluster $2a$ composed of 15,403 human accounts. Then we merged the two clusters and obtained a dataset of accounts through which we trained and tested a Random Forest classifier to differentiate between human and troll accounts. The accounts were labeled with their cluster membership, in agreement with the type of account. Since our goal was not to develop a functional classifier but to analyze the differences in terms of feature importance, we used all 32,842 accounts to train the classifier. Testing the trained classifier with the same dataset resulted in an F1 score of 0.99. However, rather than its efficacy, our interest turned to the important features on which the classifier's choices relied, also readable as those that better distinguish trolls of cluster $1a$ from humans of cluster $2a$. Indeed, we carried out a feature importance analysis using the SHapley Additive exPlanations (SHAP) values [50]. SHAP values show how each feature contributes to the prediction of an instance, either positively or negatively. These values can be exploited to determine the features that better differentiate the two clusters. In Fig. 6 are reported the ten most important features for the classifier obtained through SHAP: for each feature is reported its value, expressed by color (red corresponds to high values, while blue corresponds to low values) and the impact it has on the prediction of the classifier, which can be positive if it led to cluster $1a$, or negative if it led instead to cluster $2a$. The analysis reveals that troll accounts in cluster $1a$ retweet a wider variety of accounts (Mean retweeted accounts), and share more URLs (Tweet-URL ratio) than humans in cluster $2a$ do, which may point to an attempt to boost the account's positive perception [51]. The hypothesis on the presence of bias is disproved since the marked distinction between trolls and humans in our experiments is induced by features that have already been observed in previous work, where both high URLs sharing and anomalies in retweeting patterns have already been highlighted [13,36].

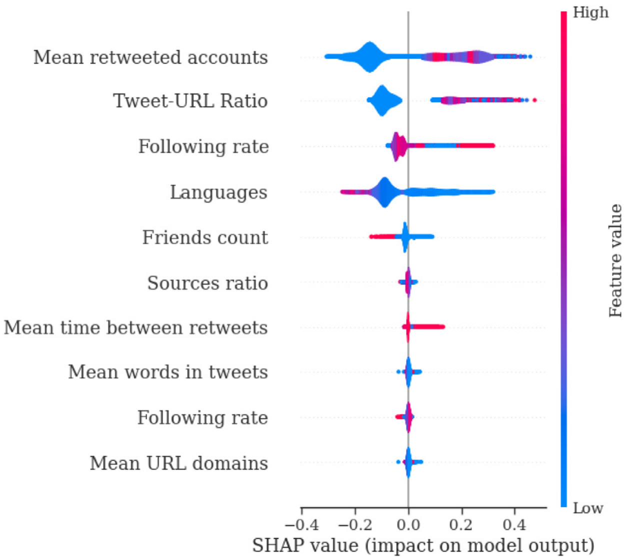


Fig. 6. Features are ranked by importance from top to bottom. The x-axis shows the SHAP value of each feature. Positive SHAP values indicate that the feature influences the classifier's prediction towards 1 (cluster 1a), and negative SHAP values suggest that the feature drives the prediction to 0 (cluster 2a). The feature value is indicated by color: red means that the feature value is high, and blue means low. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.5. Social bots distribution

Since many of the smaller clusters are populated by social bots, we analyzed these clusters composed of more than half of social bots to understand better why they are distributed in such a heterogeneous manner. Table 4 reports the clusters containing mainly social bots and, for each of them, the dominant origin datasets of the contained social bots. We observe that most of these clusters are made up of social bots from the same dataset. This distribution reflects the resilience of social bot detection models [52], which suffer from recall degradation in cross-domain evaluation [20,28]. This is due to the implicit nature of the datasets built by the research community, including those used in this work, which were collected to develop detection models for particular subtypes of social bots, thus often containing biases [52]. Furthermore, we observe that the social bots from the dataset `cresci-stock-2018` spread among various clusters; this is explained by the

Table 4
List of dominant datasets in clusters identified as social bots clusters.

Cluster	Dominant dataset	N. Accounts	%
	fake_followers-2017	759	19.33
4	traditional_spambots-2017	483	13.55
	cresci-rtbust-2019	204	5.19
5	social_spambots_2–2017	3391	99.85
6	cresci-stock-2018	1.318	40.34
7	cresci-stock-2018	3,134	98.99
8	cresci-stock-2018	3,153	99.87
9	cresci-stock-2018	2,089	77.80
11	cresci-stock-2018	1,484	70.70
12	cresci-stock-2018	1,497	82.34
14	pronbots-2019	1,194	78.30
15	cresci-stock-2018	1,008	66.62
	botwiki-2019	607	48.17
16	socialspambots3–2015	427	33.89
	gilani-2017	153	12.14
20	cresci-stock-2018	1,036	96.19
21	cresci-stock-2018	1,017	99.51
22	fake_followers-2017	1,010	99.90
24	socialspambots3–2015	821	98.68
25	cresci-stock-2018	731	100.00
26	cresci-stock-2018	712	99.72
27	cresci-stock-2018	314	45.64
32	traditional_spambots-2017	249	66.94
33	cresci-stock-2018	336	99.40

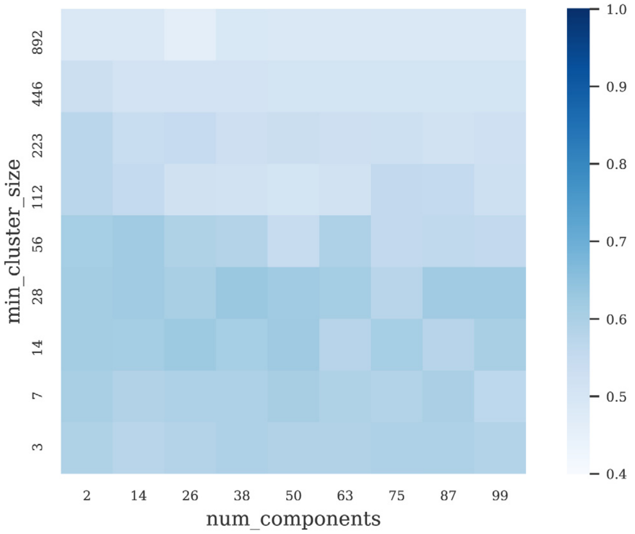


Fig. 7. Heatmap reporting the purity values obtained using the combination of traits and the origin dataset as label. Darker shades correspond to higher purity values.

nature of the dataset, which was built with accounts presenting coordination in their activity [26], making them unsuitable for feature-based distinction methods.

5.6. Datasets analysis

Starting from the results obtained with the combination of the three traits, we calculated the purity in relation to the origin dataset for each account. Fig. 7 shows the results, highlighting that although the accounts belonging to the same cluster are mainly of the same category, they do not originate from the same dataset. This result, together with the one shown previously in Fig. 1(d) and the one observed for social bots, suggests that while for social bots, the distribution among the various clusters is influenced by the dataset of origin, for trolls and humans, this is not true. However, they are reported in the same dataset, and it seems that they do not necessarily manifest similarities in terms of credibility, initiative, and adaptability.

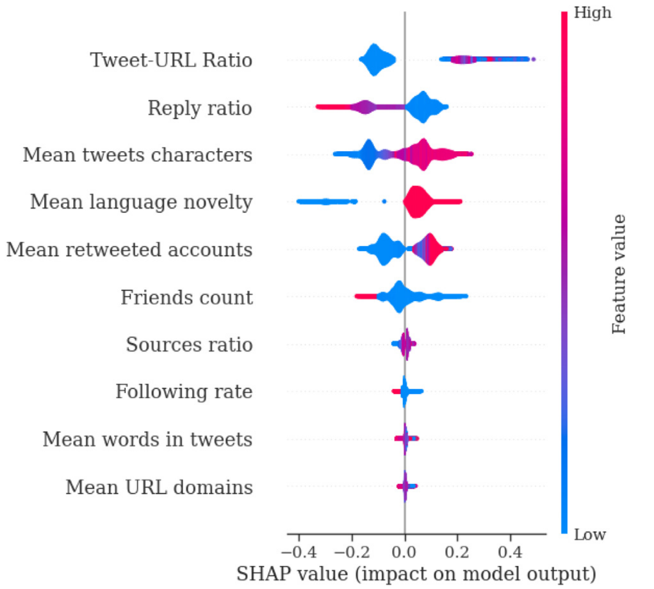


Fig. 8. Features are ranked by importance from top to bottom. The x-axis shows the SHAP value of each feature. Positive SHAP values indicate that the feature influences the classifier's prediction towards 1 (troll), and negative SHAP values suggest that the feature drives the prediction to 0 (bot). The feature value is indicated by color: red means that the feature value is high, and blue means low. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

6. Direct comparisons

So far, we have analyzed accounts using an unsupervised approach that has allowed us to characterize the different categories of accounts. State-backed trolls tend to be grouped homogeneously: about 83% of them are located in one of the two largest clusters. Social bots, in contrast, are distributed among many clusters mainly due to their origin dataset. Humans are in the middle; about 55% are found in one cluster, while the remaining are placed in several clusters along with social bots. We finalize our work with a features importance analysis analogous to the one performed in Section 5.4. This time we encompass all the accounts and compare in pairs the different categories to find out which features best distinguish them. We trained a binary classifier for every pair of different actor types and identified the most important features contributing to differentiation using SHAP. To train the classifiers, we used all accounts belonging to the two categories examined, obtaining in all cases an F1 score of 0.99, giving as input the same accounts.

6.1. Trolls vs. humans

We used the 31,013 troll accounts and 27,877 human accounts available in the dataset to train a binary classifier to bring out the most important features that differentiate state-backed trolls from humans. Applying SHAP, we obtain the results shown in Fig. 8. As already observed in Fig. 6, both the presence of URLs in tweets (Tweet-URL ratio) and the number of retweeted accounts (Mean retweeted accounts) play a key role in distinguishing these accounts categories. Moreover, it appears that the number of replies in relation to the account's overall content production (Reply ratio) affects too: humans have a higher reply rate when compared to trolls. Furthermore, two other traits emerge in the distinction between humans from trolls. According to the SHAP analysis, troll accounts are inclined to use more characters (Mean tweets characters) and introduce new words in their tweets (Mean language novelty).

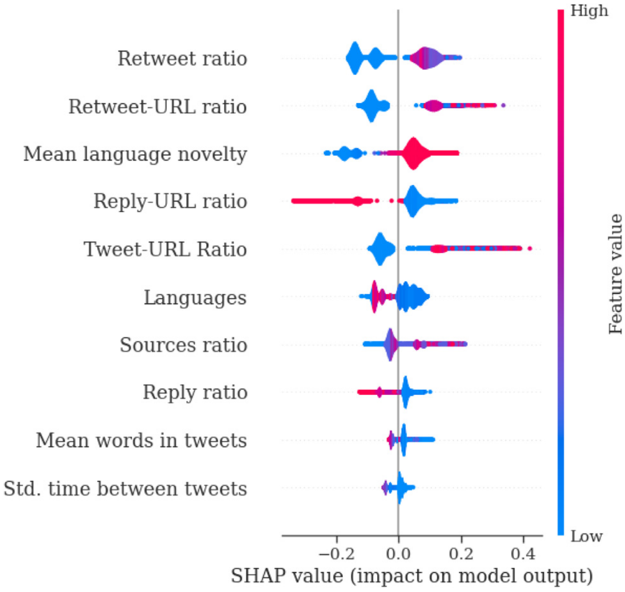


Fig. 9. Features are ranked by importance from top to bottom. The x-axis shows the SHAP value of each feature. Positive SHAP values indicate that the feature influences the classifier's prediction towards 1 (troll), and negative SHAP values suggest that the feature drives the prediction to 0 (human). The feature value is indicated by color: red means the feature value is high, and blue means low. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

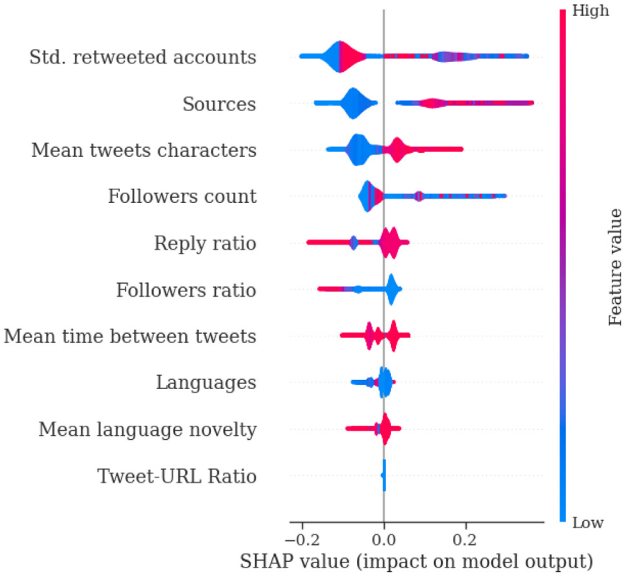


Fig. 10. Features are ranked by importance from top to bottom. The x-axis shows the SHAP value of each feature. Positive SHAP values indicate that the feature influences the classifier's prediction towards 1 (bot), and negative SHAP values suggest that the feature drives the prediction to 0 (human). The feature value is indicated by color: red means that the feature value is high, and blue means low. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

6.2. Trolls vs. social bots

When using the 31,013 troll accounts and 32,742 social bot accounts to train a binary classifier, through SHAP analysis, we obtain the results shown in Fig. 9. Compared to social bots, troll accounts produce more retweets (Retweet ratio), and such retweets contain more URLs (Retweet-URL ratio). Even their tweets contain more URLs (Tweet-URL Ratio) and, in contrast to social bots, which are automated accounts, they produce content with a higher language novelty (Mean language novelty). As for replies, social bots show a higher ratio in terms of overall production (Reply ratio) and include more URLs in their replies (Reply-URL ratio).

6.3. Social bots vs. humans

The last comparison involves the 32,742 social bot accounts and the 27,877 human accounts. This topic has been significantly discussed in the literature over the past decade [7]. Results reported in Fig. 10 are consistent with the challenge of discriminating between the social bot and human accounts reported in the literature. In fact, in this instance, only two features report a clear division with respect to the classifier decision and value. Beyond having more characters (Mean tweets characters), tweets produced by social bots are generated from multiple sources (Sources): this is quite evident since social bots often use Twitter APIs to automatically share tweets, while humans usually rely on a limited number of interfaces available for desktop and mobile devices, e.g., Twitter for Android.

7. Conclusions and future work

In this work, we investigated accounts belonging to different datasets in order to characterize the different accounts' categories. To this end, our work provides several contributions. First, we proposed three account traits, each represented by a set of features. We showed that the traits we identified can distinguish different actors with different results. In particular, we obtained our best results by combining them. Second, we analyzed our results qualitatively — by visualizing and studying how the accounts place in the bidimensional space according to the traits identified, and quantitatively — by leveraging well-known metrics such as cluster purity and investigating the obtained account distributions. We also cross-checked our clustering results with Twitter's categorization of IOs and the different datasets considered in this work. Finally, we performed a feature importance analysis highlighting the features that help differentiate the different actors when taken in pairs. In future work, we want to test the effectiveness of our traits on datasets where the accounts operated in the same scenario. Our analysis revealed that troll accounts retweet a wider variety of accounts and share more URLs than humans, while social bots appear to be defined by their origin dataset.

Moreover, while for the distinction between trolls and the other actors, several features emerged, human and social bot accounts only present two features that clearly separate them. The results of our work can contribute to understanding the characteristics of the different types of accounts on Twitter. However, the same analysis can be applied to datasets from other social networks, such as Facebook and Instagram. In addition, our analysis increases the knowledge of trolls, which so far have received limited attention from quantitative researchers. Finally, our work also paves the way for deploying computational tools to assist analysts with the cumbersome manual investigations required to make sense of large malicious campaigns. We expect our endeavor to foster future multidisciplinary studies, providing journalists, social and political analysts, policymakers, and other stakeholders with computational tools to navigate the increasingly complex online information landscape.

CRediT authorship contribution statement

Michele Mazza: Conceptualization, Methodology, Investigation, Software, Writing – original draft. **Marco Avvenuti:** Writing – review & editing, Supervision. **Stefano Cresci:** Investigation, Writing – review & editing, Supervision. **Maurizio Tesconi:** Conceptualization, Writing – review & editing, Supervision, Funding acquisition.

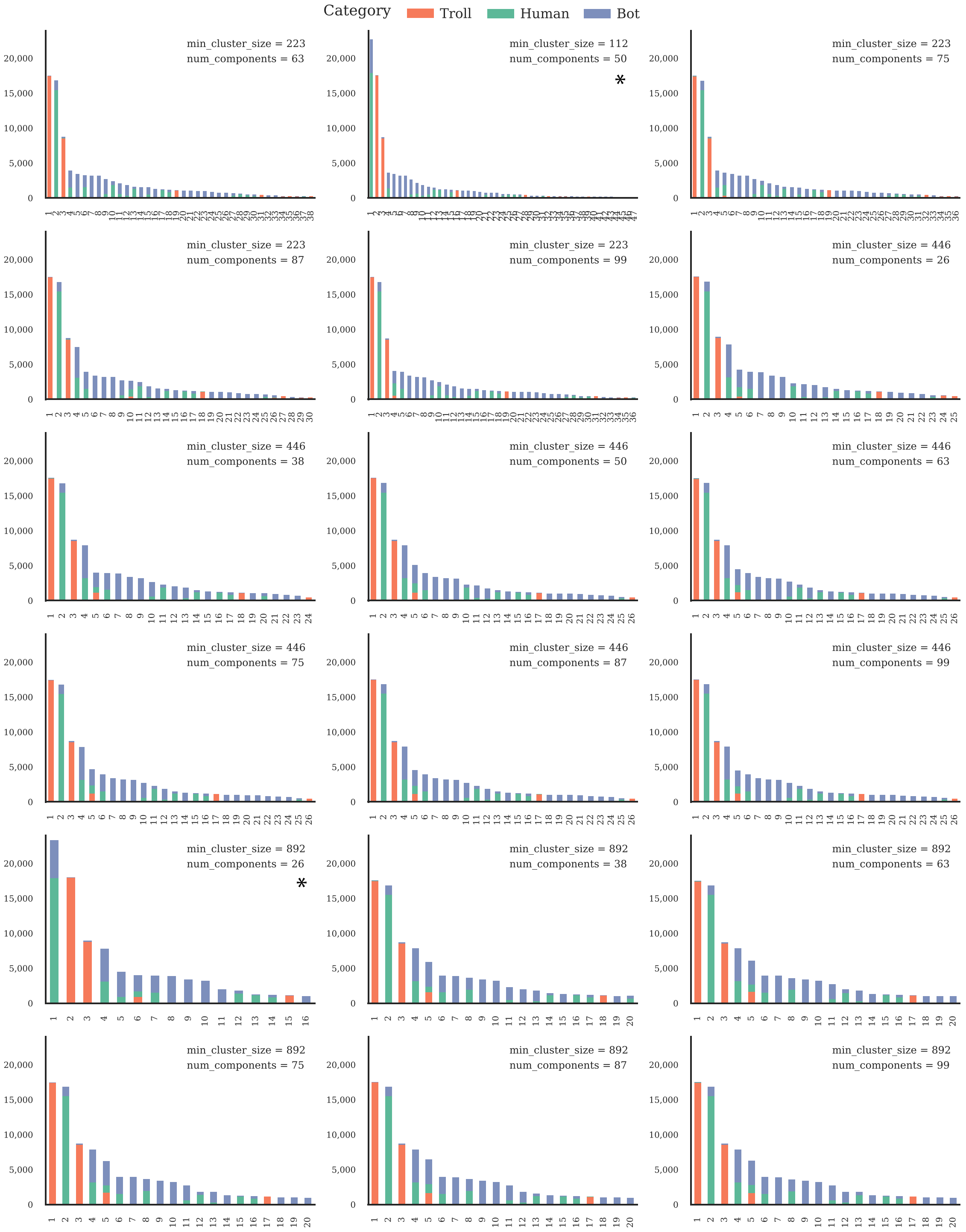


Fig. A.11. Purity values obtained using the combination of traits and the origin dataset as label.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The dataset can be build through instructions present on the paper.

Acknowledgments

This work was supported by the EU H2020 Program under the scheme INFRAIA-01-2018-2019 Research and Innovation action grant agreement #871042 SoBigData++ European Integrated Infrastructure for Social Mining and Big Data Analytics and by the Italian Ministry of Education and Research (MIUR) in the framework of the CrossLab project (Departments of Excellence).

Appendix. Cluster distribution of experimental setups

Fig. A.11 shows the distribution of the three different categories of accounts over the clusters obtained in experimental setups, presenting less than 50 clusters and a silhouette coefficient greater than 0.7

References

[1] C. Wardle, H. Derakhshan, Information disorder: Toward an interdisciplinary framework for research and policy making, Counc. Eur. Rep. 27 (2017) 1–107.

[2] S. Bradshaw, P.N. Howard, Challenging truth and trust: A global inventory of organized social media manipulation, Comput. Propag. Proj. 1 (2018).

[3] R. DiResta, K. Shaffer, B. Ruppel, D. Sullivan, R. Matney, R. Fox, J. Albright, B. Johnson, The tactics & tropes of the internet research agency, New Knowl. (2018) URL https://www.yonder-ai.com/resources/the-disinformation-report/.

[4] L. Nizzoli, S. Tardelli, M. Avvenuti, S. Cresci, M. Tesconi, Coordinated behavior on social media in 2019 UK general election, in: Proceedings of the International AAAI Conference on Web and Social Media, Vol. 15, 2021, pp. 443–454.

[5] M. Cinelli, S. Cresci, W. Quattrociocchi, M. Tesconi, P. Zola, Coordinated inauthentic behavior and information spreading on Twitter, Decis. Support Syst. (2022) 1–12.

[6] M. Mazza, G. Cola, M. Tesconi, Ready-to-(ab)use: From fake account trafficking to coordinated inauthentic behavior on Twitter, Online Soc. Netw. Media 31 (2022) 100224, http://dx.doi.org/10.1016/j.osnem.2022.100224, URL https://www.sciencedirect.com/science/article/pii/S2468696422000271.

[7] S. Cresci, A decade of social bot detection, Commun. ACM 63 (10) (2020) 72–83, http://dx.doi.org/10.1145/3409116.

[8] K. Starbird, Disinformation's spread: bots, trolls and all of us, Nature 571 (7766) (2019) 449–450.

[9] A. Trujillo, S. Cresci, Make reddit great again: Assessing community effects of moderation interventions on r/the_Donald, in: The 25th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW'22), ACM, 2022.

[10] S. Cresci, A. Trujillo, T. Fagni, Personalized interventions for online moderation, in: The 33rd ACM Conference on Hypertext and Social Media (HT'22), ACM, 2022, pp. 248–251.

[11] K. Hristakieva, S. Cresci, G. Da San Martino, M. Conti, P. Nakov, The spread of propaganda by coordinated communities on social media, in: The 14th ACM Web Science Conference (WebSci'22), 2022, pp. 191–201.

[12] P.M. Barrett, Disinformation and the 2020 election: how the social media industry should prepare, NYU Stern Cent. Bus. Hum. Rights 1 (2019).

[13] S. Zannettou, T. Caulfield, E. De Cristofaro, M. Sirivianos, G. Stringhini, J. Blackburn, Disinformation warfare: Understanding state-sponsored trolls on Twitter and their influence on the web, in: Companion Proceedings of the 2019 World Wide Web Conference, WWW '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 218–226, http://dx.doi.org/10.1145/3308560.3316495.

[14] O. Varol, E. Ferrara, C. Davis, F. Menczer, A. Flammini, Online human-bot interactions: Detection, estimation, and characterization, in: Proceedings of the International AAAI Conference on Web and Social Media, 11, (1) 2017, pp. 280–289, URL https://ojs.aaai.org/index.php/ICWSM/article/view/14871.

[15] O. Varol, C.A. Davis, F. Menczer, A. Flammini, Feature engineering for social bot detection, in: G. Dong, H. Liu (Eds.), Feature Engineering for Machine Learning and Data Analytics, in: Data Mining and Knowledge Discovery Series, Chapman and Hall/CRC Press, 2018, pp. 311–334, URL https://www.crcpress.com/Feature-Engineering-for-Machine-Learning-and-Data-Analytics/Dong-Liu/p/book/9781138744387.

[16] J. Im, E. Chandrasekharan, J. Sargent, P. Lighthammer, T. Denby, A. Bhargava, L. Hemphill, D. Jurgens, E. Gilbert, Still out there: Modeling and identifying Russian troll accounts on Twitter, in: 12th ACM Conference on Web Science, WebSci '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 1–10, http://dx.doi.org/10.1145/3394231.3397889.

[17] K. Starbird, A. Arif, T. Wilson, Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations, Proc. ACM Hum.-Comput. Interact. 3 (CSCW) (2019) http://dx.doi.org/10.1145/3359229.

[18] M. Mazza, S. Cresci, M. Avvenuti, W. Quattrociocchi, M. Tesconi, Rtbust: Exploiting temporal patterns for botnet detection on Twitter, in: Proceedings of the 10th ACM Conference on Web Science, Association for Computing Machinery, New York, NY, USA, 2019, pp. 183–192, http://dx.doi.org/10.1145/3292522.3326015.

[19] S. Liu, B. Hooi, C. Faloutsos, HoloScope: Topology-and-spike aware fraud detection, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM '17, Association for Computing Machinery, New York, NY, USA, 2017, pp. 1539–1548, http://dx.doi.org/10.1145/3132847.3133018.

[20] K.-C. Yang, O. Varol, P.-M. Hui, F. Menczer, Scalable and generalizable social bot detection through data selection, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, (01) 2020, pp. 1096–1103.

[21] B. Ghanem, D. Buscaldi, P. Rosso, TexTrolls: Identifying Russian trolls on Twitter from a textual perspective, 2019, arXiv preprint arXiv:1910.01340.

[22] L. Luceri, S. Giordano, E. Ferrara, Detecting troll behavior via inverse reinforcement learning: A case study of Russian trolls in the 2016 US election, in: Proceedings of the International AAAI Conference on Web and Social Media, 14, (1) 2020, pp. 417–427, URL https://ojs.aaai.org/index.php/ICWSM/article/view/7311.

[23] I. Alsmadi, M.J. O'Brien, How many bots in Russian troll tweets? Inf. Process. Manage. 57 (6) (2020) 102303, http://dx.doi.org/10.1016/j.ipm.2020.102303, URL https://www.sciencedirect.com/science/article/pii/S0306457320307986.

[24] L. Luceri, F. Cardoso, S. Giordano, Down the bot hole: Actionable insights from a one-year analysis of bot activity on Twitter, First Monday 26 (3) (2021) http://dx.doi.org/10.5210/fm.v26i3.11441.

[25] A. Bessi, E. Ferrara, Social bots distort the 2016 US presidential election online discussion, First Monday 21 (11–7) (2016).

[26] S. Cresci, F. Lillo, D. Regoli, S. Tardelli, M. Tesconi, Cashtag piggybacking: Uncovering spam and bot activity in stock microblogs on Twitter, ACM Trans. Web 13 (2) (2019) http://dx.doi.org/10.1145/3313184.

[27] S. Kudugunta, E. Ferrara, Deep neural networks for bot detection, Inform. Sci. 467 (2018) 312–322.

[28] M. Sayyadiharikandeh, O. Varol, K.-C. Yang, A. Flammini, F. Menczer, Detection of novel social bots by ensembles of specialized classifiers, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 2725–2732, http://dx.doi.org/10.1145/3340531.3412698.

[29] S. Tardelli, M. Avvenuti, M. Tesconi, S. Cresci, Detecting inorganic financial campaigns on Twitter, Inf. Syst. 103 (2022) 101769.

[30] P. Zola, G. Cola, M. Mazza, M. Tesconi, Interaction strength analysis to model retweet cascade graphs, Appl. Sci. 10 (23) (2020) http://dx.doi.org/10.3390/app10238394, URL https://www.mdpi.com/2076-3417/10/23/8394.

[31] K.-C. Yang, O. Varol, C.A. Davis, E. Ferrara, A. Flammini, F. Menczer, Arming the public with artificial intelligence to counter social bots, Hum. Behav. Emerg. Technol. 1 (1) (2019) 48–61, http://dx.doi.org/10.1002/hbe2.115, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbe2.115, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hbe2.115.

[32] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, M. Tesconi, Emergent properties, models and laws of behavioral similarities within groups of Twitter users, Comput. Commun. 150 (2020) 47–61.

[33] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, M. Tesconi, 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, in: WWW '17 Companion, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2017, pp. 963–972, http://dx.doi.org/10.1145/3041021.3055135.

[34] M. Bastos, D. Mercea, The public accountability of social platforms: lessons from a study on bots and trolls in the brexit campaign, Phil. Trans. R. Soc. A 376 (2128) (2018) 20180003, http://dx.doi.org/10.1098/rsta.2018.0003, URL https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2018.0003.

[35] A. Addawood, A. Badawy, K. Lerman, E. Ferrara, Linguistic cues to deception: Identifying political trolls on social media, in: Proceedings of the International AAAI Conference on Web and Social Media, 13, (01) 2019, pp. 15–25, URL https://ojs.aaai.org/index.php/ICWSM/article/view/3205.

[36] S. Zannettou, T. Caulfield, W. Setzer, M. Sirivianos, G. Stringhini, J. Blackburn, Who let the trolls out? Towards understanding state-sponsored trolls, in: Proceedings of the 10th ACM Conference on Web Science, Association for Computing Machinery, New York, NY, USA, 2019, pp. 353–362, http://dx.doi.org/10.1145/3292522.3326016.

[37] R.L. Boyd, A. Spangher, A. Fourney, B. Nushi, G. Ranade, J. Pennebaker, E. Horvitz, Characterizing the internet research agency's social media operations during the 2016 US presidential election using linguistic analyses, 2018.

[38] A. Atanasov, G. De Francisci Morales, P. Nakov, Predicting the role of political trolls in social media, in: Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 1023–1034, http://dx.doi.org/10.18653/v1/K19-1096, URL https://aclanthology.org/K19-1096.

[39] D.L. Linvill, P.L. Warren, Troll factories: Manufacturing specialized disinformation on Twitter, Political Commun. 37 (4) (2020) 447–467, http://dx.doi.org/10.1080/10584609.2020.1718257.

[40] D. Kim, T. Graham, Z. Wan, M.-A. Rizoiu, Analysing user identity via time-sensitive semantic edit distance (t-SED): A case study of Russian trolls on Twitter, J. Comput. Soc. Sci. 2 (2) (2019) 331–351, http://dx.doi.org/10.1007/s42001-019-00051-x.

[41] Z. Gilani, R. Farahbakhsh, G. Tyson, L. Wang, J. Crowcroft, Of bots and humans (on Twitter), in: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, ASONAM '17, Association for Computing Machinery, New York, NY, USA, 2017, pp. 349–354, http://dx.doi.org/10.1145/3110025.3110090.

[42] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, M. Tesconi, Fame for sale: Efficient detection of fake Twitter followers, Decis. Support Syst. 80 (2015) 56–71, http://dx.doi.org/10.1016/j.dss.2015.09.003, URL https://www.sciencedirect.com/science/article/pii/S0167923615001803.

[43] K. Beyer, J. Goldstein, R. Ramakrishnan, U. Shaft, When is ''nearest neighbor'' meaningful? in: C. Beeri, P. Buneman (Eds.), Database Theory — ICDT'99, Springer Berlin Heidelberg, Berlin, Heidelberg, 1999, pp. 217–235.

[44] L. McInnes, J. Healy, N. Saul, L.G. berger, UMAP: Uniform manifold approximation and projection, J. Open Source Softw. 3 (29) (2018) 861, http://dx.doi.org/10.21105/joss.00861.

[45] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (86) (2008) 2579–2605, URL http://jmlr.org/papers/v9/vandermaaten08a.html.

[46] R.J.G.B. Campello, D. Moulavi, J. Sander, Density-based clustering based on hierarchical density estimates, in: J. Pei, V.S. Tseng, L. Cao, H. Motoda, G. Xu (Eds.), Advances in Knowledge Discovery and Data Mining, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 160–172.

[47] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al., A density-based algorithm for discovering clusters in large spatial databases with noise, in: KDD, Vol. 96, (34) 1996, pp. 226–231.

[48] B.E. Dom, An information-theoretic external cluster-validity measure, in: Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence, UAI '02, Morgan Kaufmann Publishers Inc., 2002, pp. 137–145.

[49] M. Kim, R.S. Ramakrishna, New indices for cluster validity assessment, Pattern Recognit. Lett. 26 (15) (2005) 2353–2363, http://dx.doi.org/10.1016/j.patrec.2005.04.007.

[50] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, NIPS '17, Curran Associates Inc., Red Hook, NY, USA, 2017, pp. 4768–4777.

[51] A. Gupta, P. Kumaraguru, Credibility ranking of tweets during high impact events, in: Proceedings of the 1st Workshop on Privacy and Security in Online Social Media, PSOSM '12, Association for Computing Machinery, New York, NY, USA, 2012, http://dx.doi.org/10.1145/2185354.2185356.

[52] J. Echeverrïa, E. De Cristofaro, N. Kourtellis, I. Leontiadis, G. Stringhini, S. Zhou, LOBO: Evaluation of generalization deficiencies in Twitter bot classifiers, in: Proceedings of the 34th Annual Computer Security Applications Conference, ACSAC '18, Association for Computing Machinery, New York, NY, USA, 2018, pp. 137–146, http://dx.doi.org/10.1145/3274694.3274738.