REGULAR ARTICLE   **Open Access**

Exposing influence campaigns in the age of LLMs: a behavioral-based AI approach to detecting state-sponsored trolls

Fatima Ezzeddine[1,2]* , Omran Ayoub[1], Silvia Giordano[1], Gianluca Nogara[1], Ihab Sbeity[2], Emilio Ferrara[3] and Luca Luceri[1,3]

*Correspondence:
fatima.ezzeddine@supsi.ch
[1]Department of Innovative Technologies, University of Applied Sciences and Arts of Southern Switzerland, Lugano, Switzerland
[2]Department of Applied Mathematics, Faculty of Science, Lebanese University, Beirut, Lebanon
Full list of author information is available at the end of the article

Abstract

The detection of state-sponsored trolls operating in influence campaigns on social media is a critical and unsolved challenge for the research community, which has significant implications beyond the online realm. To address this challenge, we propose a new AI-based solution that identifies troll accounts solely through behavioral cues associated with their sequences of sharing activity, encompassing both their actions and the feedback they receive from others. Our approach does not incorporate any textual content shared and consists of two steps: First, we leverage an LSTM-based classifier to determine whether account sequences belong to a state-sponsored troll or an organic, legitimate user. Second, we employ the classified sequences to calculate a metric named the "Troll Score", quantifying the degree to which an account exhibits troll-like behavior. To assess the effectiveness of our method, we examine its performance in the context of the 2016 Russian interference campaign during the U.S. Presidential election. Our experiments yield compelling results, demonstrating that our approach can identify account sequences with an AUC close to 99% and accurately differentiate between Russian trolls and organic users with an AUC of 91%. Notably, our behavioral-based approach holds a significant advantage in the ever-evolving landscape, where textual and linguistic properties can be easily mimicked by Large Language Models (LLMs): In contrast to existing language-based techniques, it relies on more challenging-to-replicate behavioral cues, ensuring greater resilience in identifying influence campaigns, especially given the potential increase in the usage of LLMs for generating inauthentic content. Finally, we assessed the generalizability of our solution to various entities driving different information operations and found promising results that will guide future research.

Keywords: Social network; Troll; Misinformation

1 Introduction

Social Media Networks (SMNs) are a crucial constituent of societies, providing a primary platform for individuals to engage in social and political discourse, as well as to disseminate critical messages and promote propaganda. SMNs have undergone a significant transformation, evolving from a simple aggregation medium to a complex ecosystem where the

Springer

line between offline and online realms is often blurred [1]. Recent studies have shown that the impact of discussions on SMNs extends beyond the online platform and can have a significant effect on societies, such as undermining the integrity of political elections and public health [2–6].

In this context, the accuracy, confidentiality, and authenticity of shared content are crucial elements for safe communication and, therefore, the well-being of societies. However, SMNs have experienced a shortage of these elements, as their growth has led to an increase in deceptive and fraudulent accounts that intentionally damage the credibility of online discussions [7]. The activity of these accounts often results in online harms that threaten the honesty and ethics of conversations, such as the propagation of hate speech, incitement of violence, and dissemination of misleading and controversial content. This has been observed in recent debates concerning the Ukraine-Russia conflict [8], Covid-19 pandemic [9–13], as well as the rise of conspiracy theories [14–16]. These fraudulent accounts represent a significant threat to healthy online conversations, whose activity has the potential to exacerbate societal divisions and affect the sovereignty of elections [17–22].

In the political sphere, Russian meddling in the 2016 U.S. Presidential election represents the most prominent case of deceptive online interference campaign [23, 24]. The Mueller report [25] suggests that Russia engaged in extensive attacks on the U.S. election system to manipulate the outcome of the 2016 voting event. The "sweeping and systematic" interference allegedly used bots (i.e., automated accounts) and trolls (i.e., state-sponsored human operators) to spread politically biased and false information [26]. In the aftermath of the election, the U.S. Congress released a list of 2752 Twitter accounts associated with Russia's "Internet Research Agency" (IRA), known as Russian trolls. As a result, significant research efforts were launched to identify fraudulent accounts and deceptive activity on several SMNs. Among these platforms, Twitter has been continuously working to eliminate malicious entities involved in information operations across different countries [27–29] and different geopolitical events [30, 31]. While there are several proven techniques for uncovering bot accounts [32–38], the detection of troll accounts is currently an unsolved issue for the research community, due to several factors tied with the human character of trolls [39]. Note that throughout this manuscript, our definition of *troll* is limited to state-sponsored human actors who have a political agenda and operate in coordinated influence campaigns, disregarding thus other hateful and harassing online activities tied with Internet-mediated trolling behavior.

Recent efforts have devised approaches for identifying trolls by leveraging linguistic cues and profile meta-data [40–44]. Although these approaches have shown promising results, they suffer from certain limitations. Some of these methods are language-dependent, focusing solely on specific spoken languages associated with the trolls under investigation [45, 46]. Others are constrained to a single SMN, relying on profile metadata and platform-specific information. Furthermore, the ease of imitating language and linguistic cues has increased with the emergence of Large Language Models (LLMs), such as ChatGPT and similar technologies. As we look ahead, our ability to detect influence operations based solely on linguistic cues may be hindered by the growing reliance on LLMs for such operations [47, 48]. These significant limitations have prompted research efforts to develop language- and content-agnostic approaches, as demonstrated in the work of Luceri et al. [49]. This approach distinguishes troll accounts by uncovering behavioral incentives from their observed activities using an Inverse Reinforcement Learning (IRL) framework.

Given that mimicking behaviors and incentives is notably more challenging than imitating language, incorporating behavioral cues either in addition to or as an alternative to purely linguistic-based methods emerges as a promising strategy in an uncertain future, particularly when the cost of generating inauthentic, yet credible, content appears to be exceptionally low [50, 51].

In this work, we advance along this research line and propose a novel approach to identify state-sponsored troll activity solely based on behavioral cues linked to accounts' sharing activities on Twitter. Specifically, we consider online activities regardless of the content shared, the language used, and the linked metadata to classify accounts as trolls or organic, legitimate users (from now on, simply *users*). Our approach aims to capture cues of behavior that differentiate trolls from users by analyzing their interactions and responses to feedback. For this purpose, we consider both the actions performed by an account, namely *active online activities*, and the feedback received by others, namely *passive online activities*, e.g., received replies and retweets. We refer to the sequence of active and passive activities as a *trajectory*, in accordance with [49]. We demonstrate the validity of our approach by detecting Russian trolls involved in the interference of the 2016 U.S. Presidential election. We also evaluate whether the proposed approach can be effectively used to identify various entities involved in diverse Twitter information operations during the 2020 U.S. Presidential election.

Contributions of this work   The core contributions of this work are summarized as follows:

- We propose a novel approach based on Long Short-Term Memory (LSTM) for classifying accounts' trajectories. Our approach correctly identifies trolls' and users' trajectories with an AUC and an F1-score of about 99%.
- Leveraging the classified trajectories, we introduce a metric, namely the *Troll Score*, that enables us to quantitatively assess the extent to which an account exhibits behavior akin to that of a state-sponsored troll. We propose a *Troll Score*-based classifier that can effectively detect troll accounts with remarkable accuracy, achieving an AUC of about 91% (F1-score ∼90%). Our approach outperforms existing behavioral-based methods and approaches the classification performance of existing linguistic solutions, all while not requiring access to the content of shared messages. This feature enhances its robustness, especially given the possibility of increased usage of LLMs for influence operations.
- By analyzing the active and passive activities in which accounts engage, we uncovered three distinct, naturally emerging *behavioral* clusters where trolls intermingle with user accounts. This finding confirms the difficulty of differentiating these two account classes when their trajectories are not considered.
- We demonstrate the capability of our approach to generalize and accurately identify diverse actors responsible for driving information operations. The results reveal that our methodology achieves an AUC of 80% (F1-score ∼82%) in detecting the drivers of different campaigns, indicating promising results for its applicability across countries, languages, and various malicious entities.

2 Related work

In this Section, we survey research on the automated detection of malicious accounts operated by trolls, with a focus on the troll farm connected to the IRA [52]. Some of these

efforts have proposed linguistic approaches that rely on the content posted by trolls to identify and detect them. For instance, [45] presented a theory-driven linguistic study of Russian trolls' language and demonstrated how deceptive linguistic signals can contribute to accurate troll identification. Similarly, [46] proposed an automated reasoning mechanism for hunting trolls on Twitter during the COVID-19 pandemic, which leverages a unique linguistic analysis based on adversarial machine learning and ambient tactical deception. In [53], the authors proposed a deep learning solution for troll detection on Reddit and analyzed the shared content using natural language processing techniques. Other works have considered fusing users' metadata and linguistic features, such as [40], which used profile description, stop word usage, language distribution, and bag of words features for detecting Russian trolls. Other approaches have relied on multimedia analysis, combining text, audio, and video analysis to detect improper material or behavior [54, 55]. For instance, [54] designed a platform for monitoring social media networks with the aim of automatically tracking malicious content by analyzing images, videos, and other media. In [55], the authors attempted to capture disinformation and trolls based on the existence of a firearm in images using the Object Detection API. A limitation of these works is their reliance on the content posted by accounts and on the extraction of linguistic features for troll identification. In contrast, our approach solely relies on the online behavior of accounts, specifically, the temporal sequence of online activities performed by a user. This presents an advantage over previous works, as it is independent of the language used or content shared and has, therefore, the potential to generalize to influence campaigns originating from diverse countries and be resilient to the use of LLMs for generating inauthentic content.

Previous studies have proposed sequence analysis approaches for identifying malicious accounts. For example, Kim et al. [56] used text and time as features to categorize trolls into subgroups based on the temporal and semantic similarity of their shared content. Luceri et al. [49] proposed a solution that only relies on the sequence of users' activity on online platforms to capture the incentives that the two classes of accounts (trolls vs. users) respond to. They detect troll accounts with a supervised learning approach fed by the incentives estimated via Inverse Reinforcement Learning (IRL). In [57], the authors proposed a model based on users' sequence of online actions to identify clusters of accounts with similar behavior. However, this approach was found to be ineffective in detecting Russian trolls, as reported in [49]. Similarly to these approaches, we propose a language- and content-agnostic method for identifying trolls based only on the sharing activities performed by the accounts on Twitter. We utilize deep learning, specifically LSTM, to classify the sequence of activities as belonging to either troll accounts or organic users. We leverage the classified sequences to quantify the extent to which an account behaves like a troll, a feature not available in earlier methods.

3 Problem formulation and trajectory definition

This section outlines the objectives of the proposed framework and elucidates the features, variables, and learning tasks integral to it. While existing methods for identifying troll activity in SMNs rely on linguistic and metadata features, our approach strives to be language- and content-agnostic. To achieve this, we rely only on behavioral cues and do not incorporate any textual or media content shared by the accounts, nor do we use their profile metadata. Consequently, our approach holds the potential for application across

various SMNs and is robust against the increasing use of LLMs and their potential role in influence campaigns [47, 48, 50].

To extract the unique online behaviors of trolls and organic users on Twitter, we extract the accounts' sequences of online activities. We consider their *active online activities*, including generating an original post (i.e., *tweet*), re-sharing another post (i.e., *retweet*), commenting an existing post (i.e., *reply*), or mentioning another user in a post (i.e., *mention*). In addition, we also propose to consider the feedback the accounts receive from other accounts, namely *passive online activities*, such as receiving a retweet (i.e., *being retweeted*), a comment (i.e., *being replied to*), or a mention (i.e., *being mentioned in a post*). By considering both the actions performed by the accounts and the feedback received, we aim to capture the distinct motivations driving trolls' activity, which may differ from those of organic users [49]. The rationale is that trolls might be motivated to pursue their agenda regardless of the feedback received from others, while organic users may be influenced by the level of endorsement they receive from the online community. For example, users might be more motivated to generate a new tweet when their content is re-shared by others, which is also viewed as a form of social endorsement [58, 59], or when they receive positive comments.

To formalize this approach, we model the SMN Twitter as a Markov Decision Process (MDP). Similarly to Luceri et al. [49], we represent Twitter as an environment constituted of multiple agents (i.e., Twitter accounts) that can perform a set of actions (i.e., *active online activities*) and receive feedback from the environment (i.e., *passive online activities*). Consistently with the IRL formulation in [49], we refer to the latter as *states*, as they represent the response of the Twitter environment, whereas we refer to the former as *actions*, as they indicate the *active online activities* that an account can perform on Twitter.

We consider four *actions* that can be performed by Twitter accounts:

- Original tweet (tw): to generate original content;
- Retweet (rt): to re-share content generated by others;
- Interact with others (in): to interact with other users via replies or mentions;
- No Action (no): to keep silent, i.e., the account does not perform any action.

For what pertains to *states*, we consider three possible feedback that Twitter accounts can receive:

- Retweeted (RT): an original tweet generated by the account is re-shared;
- Interacted with (IN): the account is involved by others via replies (i.e., comments to a tweet generated by the account) or mentions;
- No Interaction (NO): no feedback is received by the account.

Every account can move from one state to another when performing an action, and we refer to such a transition as a *state-action pair*. Note that an account can be only in one of the above-mentioned states and can perform only one action in any given state. By considering the accounts' timeline, we construct a *sequence* of state-action pairs that reconstruct the (observed) history of the account on Twitter. Overall, there exist only 11 possible combinations of state-action pairs that form a sequence of an account—the state-action pair *(*NO, no) is not considered as it does not describe any observable activity. It is also important to note that if an account does not react to the environment feedback, e.g., it does not perform *tw*, *rt* or *in*, it will be considered as silent, i.e., doing no action (*no*). Similarly, if an account keeps performing actions while not receiving any feedback, it will persist in the state *NO*.

By extracting and sorting the state-action pairs of every account in chronological order, we create a sequence of online activities describing the behavior of the observed trolls and users. Accordingly, we define the problem of classifying account sequences as a binary classification task with two classes: *troll* (positive class) and *user* (negative class). Section 5 details how we first classify these formed sequences as belonging to either trolls or users and then how we use these classified sequences to identify troll accounts. While our focus is on Twitter, this approach may be replicated on other SMNs like Facebook, which offer similar sharing activities.

4 Data

The dataset used to evaluate our approach consists of tweets generated by user accounts and by accounts identified as trolls involved in the 2016 U.S. election discussion over Twitter. Specifically, we consider the dataset described in [41, 45] collected by utilizing a set of keywords (detailed in [60]) related to the 2016 U.S. election, which encompasses 13.5M tweets generated by 2.4M accounts.

In the dataset used for this study, we only included accounts that had generated a minimum of ten active online activities and ten passive online activities. This decision was based on the findings of a previous study [49], which demonstrated that using fewer than ten active and passive online activities had a negative impact on classification accuracy. Among these accounts, we identified 342 troll accounts—selected from a list of 2752 Twitter accounts ascertained as Russian trolls by the U.S. Congress and that was publicly released during the investigation of Russian involvement in the U.S. 2016 Presidential election; and 1981 user accounts which generated 246k and 1.2M tweets, respectively.

5 Methodology

This Section details our proposed approach to identifying troll accounts. Section 5.1 provides an overview of our framework, while Sect. 5.2 describes the proposed methodology in more detail.

5.1 Overall framework

The proposed framework for the identification of troll accounts (see Fig. 1) consists of the following main steps:
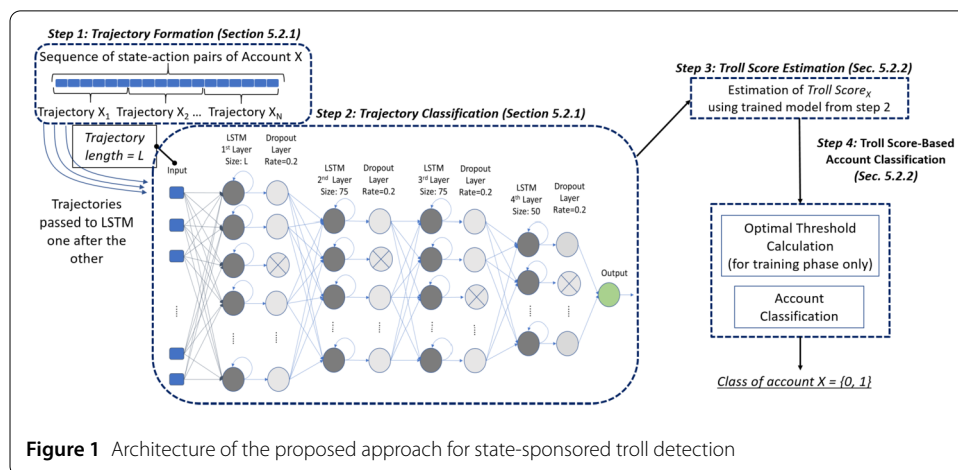

Figure 1 Architecture of the proposed approach for state-sponsored troll detection

- *Step 1—Trajectory Formation*: Creation of trajectories that represent portions of an account's sequence, where a trajectory is a time-sorted set of a pre-defined number of *state-action pairs*.
- *Step 2—LSTM-Based Trajectory Classification*: Classification of the trajectories into two classes, i.e., trolls' or users' trajectories. To perform this classification, we employ a Long Short-Term Memory (LSTM)-based classifier.
- *Step 3—Troll Score Estimation*: Computation of the *Troll Score* as the ratio of the number of trajectories classified as belonging to a troll over the number of an account's trajectories.
- *Step 4—Troll Score-Based Account Classification*: Classification of an account, i.e., *troll* or *user*, based on the computed *Troll Score.*

Figure 1 portrays an overview of our proposed framework. In *Step 1*, the sequence of online activities of an account is divided into several trajectories of state-action pairs of a pre-defined length. In *Step 2*, an LSTM model fed with the extracted trajectories classifies every trajectory as either a troll trajectory or a user trajectory. Note that, in the training phase of the LSTM model, a trajectory extracted from a sequence of a troll is considered to have a label of a *troll trajectory*, while that extracted from a sequence of a user is given the label of a *user trajectory*. We employ deep learning and, more specifically, an LSTM model as we deal with time series data (i.e., sharing activities represented by state-action trajectories). After classifying the trajectories, a Troll Score of every account is computed as explained in *Step 3*. Finally, in *Step 4*, an account is classified as either a troll or a user based on the Troll Score.

5.2 Trolls detection

We now discuss in more detail our proposed approach for the identification of troll accounts. In Sect. 5.2.1, we describe the LSTM-based model used to classify trajectories of state-action pairs, while in Sect. 5.2.2, we introduce the *Troll Score* metric and we detail the account classification approach. The code and scripts used to implement the methodological framework detailed below are freely available to the research community.[1]

5.2.1 LSTM-based model for trajectory classification

Model input. The LSTM model takes as input a trajectory of a given length, i.e., a predefined number L of state-action pairs. To build these trajectories, two questions need to be answered: *i*) How to choose the value of L?, and *ii*) given a value of L, how to divide the sequence into trajectories?

How to choose the value of L? The choice of L is not trivial, and it is, in fact, decisive. Considering a small value of L is desired from a computational standpoint. However, a relatively small value for L might not be enough for the model to distinguish troll trajectories from user trajectories. Conversely, considering a relatively large value of L might not be feasible, as not every account might largely engage in online activities, i.e., not all the accounts might have a long sequence.

Given a value of L, how to divide the sequence into trajectories? To form trajectories for trajectory classification, we consider non-overlapping parts of the sequence, i.e., we divide

---

[1]https://github.com/FatimaEzzedinee/Exposing-Influence-Campaigns-in-the-Age-of-LLMs-A-Behavioral-Based-AI-Approach.

the sequence into L-long trajectories. For instance, for $L = 100$ and a sequence composed of 200 state-action pairs, two trajectories, each of length L, are created and therefore considered for classification. Note that, for a given value of L, the number of trajectories to classify per account differs. A large L leads to a low number of trajectories to classify, while a small L leads to a high number of trajectories to classify. We will show the impact of this parameter in Sect. 6.1 by performing sensitivity analysis and monitoring the performance of our model (in terms of several classification metrics) at varying L.

5.2.2 Troll score-based classification of accounts

Troll score definition. The rationale behind defining a *Troll Score* is to have a measure to quantify the extent to which an account behaves like a troll in its online activity. We define the Troll Score of an account as the ratio between the number of trajectories classified as a troll trajectory by the LSTM model and the total number of trajectories of the account. The Troll Score, thus, ranges from zero to one, where a score close to one indicates a troll-like behavior, and a score close to zero denotes a user-like behavior. To compute the Troll Score of a given account, we consider a sliding window of length L over the whole sequence of state-action pairs of every account under scrutiny. This approach allows each state-action pair to contribute to multiple trajectories.

Troll score for account classification. To classify accounts based on their Troll Score, a threshold to distinguish the two classes is required. We compute this threshold as follows. First, the Troll Score of a subset of the accounts is computed. Then, we iterate over all threshold values ranging from 0 to 1, with a step size of 0.02, reporting the performance of each threshold value in terms of AUC. That is, we consider each value to serve as a threshold and classify accounts accordingly, where each value will result in a different classification of the accounts and, hence, a distinct performance. Finally, we select the threshold that provides the best AUC. The same method can be applied to optimize other classification metrics, e.g., precision, recall, etc. For our evaluations, we test our approach with a 10-fold cross-validation.

6 Implementation and results

In this Section, we first perform a sensitivity analysis comparing the performance of the LSTM-based trajectory classification approach in the case of trajectories composed of (*i*) state-action pairs and (*ii*) actions only. The objective here is to evaluate whether the states (i.e., received feedback) represent beneficial signals for the classification task. The rationale is to understand if the feedback received by trolls and users, along with the way they react to such stimuli, might allow us to better discern these two classes of accounts with respect to leveraging only their actions. Then, we benchmark the trajectory classification approach based on LSTM, comparing its performance to those of off-the-shelf machine learning models.

After validating the performance of the LSTM-based trajectory classification approach, we use it to compute the *Troll Score* of the accounts under investigation. We then exploit the Troll Score to distinguish troll and user accounts, and we compare the classification performance to other approaches proposed in the literature. Moreover, based on the obtained results, we conduct an observational analysis to further investigate how the visited state-action pairs of trolls and users differ. Finally, we evaluate whether the proposed approach can generalize to the detection of entities operating in influence campaigns. We

present the results of our approach on data from online discussions related to the U.S. 2020 Election on Twitter with the objective of distinguishing the drivers of influence campaigns from organic users.

6.1 Trajectory classification

This subsection presents the implementation details of our proposed LSTM model for trajectory classification and then discusses results comparing our proposed approach to off-the-shelf machine learning models.

6.1.1 Implementation details

Our proposed LSTM model for trajectory classification is composed of four LSTM layers, four dropout layers, and a dense layer. Each layer is followed by a Dropout Layer to reduce overfitting [61]. The *sigmoid* is used as an activation function for both the hidden and output layers. We fine-tune the hyper-parameters of our model with a random search. We encode trajectories to be fed into the LSTM model using *Label Encoding*, which consists of assigning an integer to each combination of state-action pairs. The entire model is trained by minimizing the binary cross-entropy with the Adam optimization algorithm [62].

6.1.2 State-action vs. action sequences

To perform trajectory classification, we rely on the historical data of the accounts. To evaluate our proposed approach, we build the trajectories of each of the accounts in two ways: (*i*) considering state-action pairs sequences or (*ii*) considering only the sequences of actions. Further, we consider five different values of the trajectory length *L*, ranging from 50 to 200, as shown in Table 1. Note that the overall number of trajectories changes with the value of *L*. We report in Table 1 the number of trajectories per every class of accounts in each of the cases (state-action pairs and only actions) and for all values of *L*. It is worth noting that the number of sequences based only on *Actions* is much lower than those based on *State-Action* pairs, as trajectories of *Actions* are formed only by three sharing activities (*tw*, *rt*, and *in*), and hence are much shorter than *State-Action* trajectories. We train and test our LSTM model in both cases and report the results of a 10-fold stratified cross-validation. It should be noted that for the scenario with only *Actions*, there is an imbalance between the number of sequences of trolls and users. To solve this issue, we employ the under-sampling technique [63, 64].

Table 1 Number of trajectories of trolls and users for each trajectory length *L*

Input	*L*	Trolls trajectories	Users trajectories
State-Action	50	64,984	102,101
	65	49,975	78,424
	80	40,587	63,643
	100	32,457	50,831
	150	21,628	33,795
	200	16,230	25,299
Action	50	7709	93,721
	65	5927	72,091
	80	4802	58,593
	100	3819	46,863
	150	2547	31,267
	200	1920	23,476

Figure 2 (a) AUC and Accuracy of the LSTM-based sequence classification approach with trajectories composed of state-action pairs or actions only, and (b) Precision, Recall, and F1-Score of the LSTM-based sequence classification approach with trajectories composed of state-action pairs or actions only
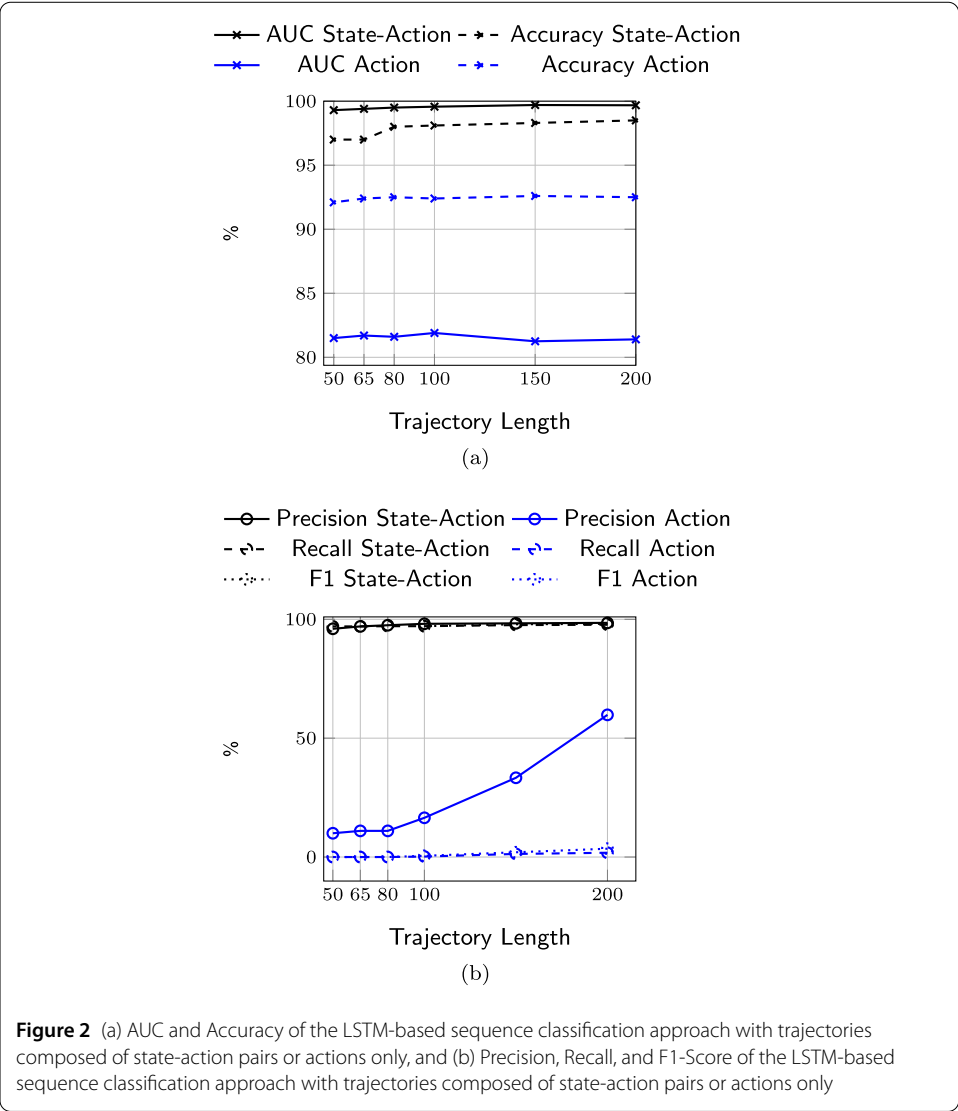
Figure 2(a) shows the AUC and accuracy of the LSTM-classifier with trajectories composed of (*i*) *State-Action* pairs and (*ii*) *Actions* only as functions of the trajectory length *L*. Results show that both classification metrics are higher when *State-Action* pairs are considered with respect to sequences of *Actions* for every value of *L*. Specifically, when sequences are composed of *State-Action* pairs, the AUC is about 99%, significantly higher than when considering *Actions* only (AUC around 82%). This is also reflected in the classification performance in terms of accuracy (97% with *State-Action* pairs vs. 92% with *Actions* only). This suggests that the *states* (i.e., feedback from other users) represent beneficial signals for the accounts classification task as, along with the *actions*, allow achieving a better distinction between trolls and users if compared to the results achieved when considering trajectories composed of actions only. From Fig. 2(a), we also note that varying the value of *L* has a minimal impact on the model's performance in both cases. In fact, the AUC of the LSTM-based approach reaches 99% even when *L* is relatively short (e.g., *L* = 50). This finding shows that the proposed model (when considering *State-Action*) has the ability to correctly identify trolls' and users' sequences even when little information

about accounts' online activity is available. This surprisingly high classification accuracy is probably due to the nature of our experimental design, which focuses on distinguishing the activity sequences of organic users from those of troll accounts. The latter, performing their agenda regardless of others' feedback [49], present activity patterns naturally different with respect to legitimate users. While these differences might appear explicit in such a high-quality dataset, we do not expect the same results in a more challenging scenario (see Sect. 6.3).
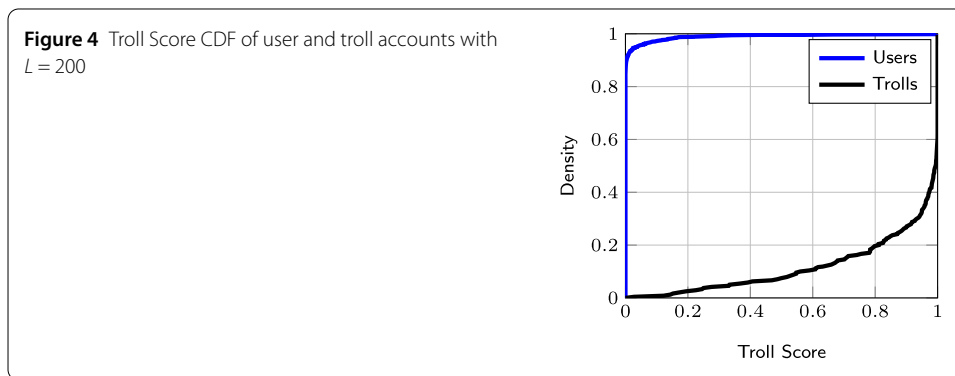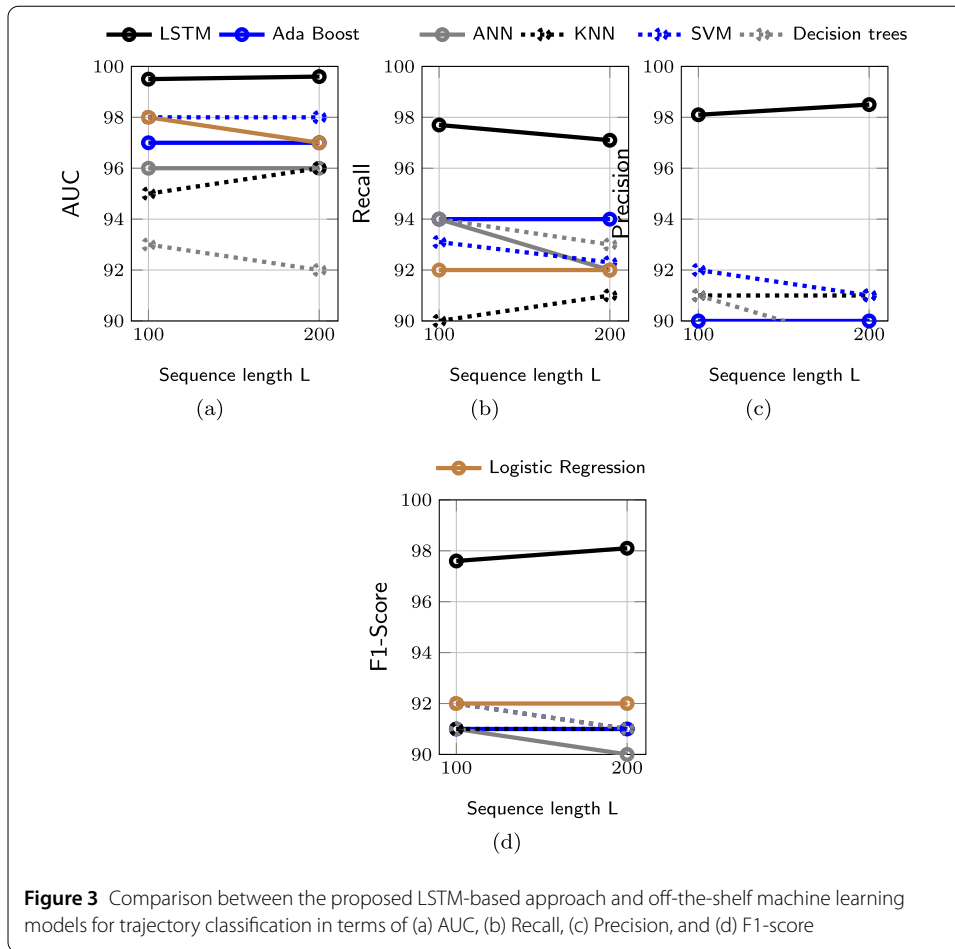
We further evaluate the performance of our model by considering other classification metrics such as Precision, Recall, and F1 score. Figure 2(b) depicts these metrics for the different values of L. Results show that the model has a nearly perfect performance with *State-Action* considering all metrics (around 98%) with a slight increase in performance as L increases. On the contrary, the LSTM classifier with *Actions* sequences suffers from low performance, nearly 0%, in terms of Recall and F1, while precision ranges between 10% and 60% and increases as L increases. This further confirms our previous intuition in complementing actions with the feedback the accounts receive from the environment (i.e., states). Indeed, these results show that the states are essential to accurately classify the two classes of accounts, which suggests that trolls might react to online feedback differently from non-trolls. In the rest of our analysis, we will consider the model based on *State-Action* pairs, given that it has been shown to be effective in discriminating between users and trolls.

6.1.3 LSTM vs. off-the-shelf machine learning models

We now compare the performance of our proposed LSTM-based model for trajectory classification to those of off-the-shelf machine learning models. Specifically, we consider Support Vector Machine (SVM), Artificial Neural Network (ANN), Decision Tree, Ada Boost, Logistic Regression, and K-Nearest Neighbors (KNN). Figure 3 shows AUC, Recall, Precision, and F1-score of these models for $L = 100$ and $L = 200$. For all metrics, and for both values of L, all machine learning models show a promising performance (e.g., AUC of 92% and higher), whereas our proposed LSTM-based approach achieves a near-optimal performance, thus outperforming all other models. More specifically, the LSTM-based approach shows 98% for Precision and Recall, significantly higher than that of other approaches, which do not exceed 92%. This indicates that trolls' and users' trajectories embed distinguishable patterns, which are more easily identifiable by models conceived to work with sequential data, such as LSTM, than other general-purpose machine learning models.

6.2 Troll score for account classification

We now focus our discussion on the classification of accounts. In this scenario, we build trajectories by sliding a window of length L with a step of one element over the entire sequence of state-action pairs. Note that in this case, two consecutive trajectories overlap by $L - 1$ elements. Indeed, the same state-action pair is considered into several trajectories, depending on its position within the sequence. As previously discussed, the classification of the accounts is based on the *Troll Score* metric. The latter is computed for every account, leveraging the classification of its trajectories extracted with a sliding window. Finally, the Troll Score of the account under scrutiny is compared to a *Troll Score threshold* (see Sect. 5.2.2) to assign the account to one of the two classes of accounts (troll vs. user).

Figure 3 Comparison between the proposed LSTM-based approach and off-the-shelf machine learning models for trajectory classification in terms of (a) AUC, (b) Recall, (c) Precision, and (d) F1-score

Figure 4 Troll Score CDF of user and troll accounts with $L = 200$

6.2.1 Troll score of user and troll accounts

We first analyze the Troll Score computed for each account in our dataset. Figure 4 shows the Cumulative Distribution Function (CDF) of the Troll Score of user and troll accounts for $L = 200$.[2] Numerical results show that the CDF of the Troll Score of users shows a logarithmic growth (the CDF reaches 0.95 for a Troll Score of 0.02). This indicates that most of the users (95% of them) have an almost-zero Troll Score while the rest (remaining

---

[2]We omit showing the case with $L = 100$ as results present same insights as for $L = 200$.

5%) have a Troll Score of at most 0.2, which can be considered considerably low. On the contrary, the CDF related to the trolls follows exponential growth: It increases very slowly for low values of Troll Score and then increases exponentially for higher values of Troll Score. In particular, results show that 80% of the trolls have a Troll Score of at least 0.8. On the one hand, this suggests that a significant portion of the trolls are characterized by a high Troll Score. On the other hand, this indicates that a limited yet considerable portion of trolls have a relatively low Troll Score, e.g., 10% have a Troll Score of at most 0.5. This finding demonstrates that some of the trolls under analysis, specifically those characterized by a low Troll Score, exhibit user-like behavior. To inspect this in more detail, we perform an analysis to observe the visited state-action pairs of all accounts. We present this observational analysis in the following section.

6.2.2 Behavioral clustering and troll score

This section presents an analysis conducted to analyze the behavior of troll and user accounts in terms of their visited state-action pairs, i.e., state and action pairs that are present in the account's trajectory. The rationale behind this analysis is to examine whether trolls and users can be distinguished by looking at their visited state-action pairs. Specifically, we evaluate whether the different classes of accounts are grouped into distinct clusters and, if clusters are observed, their relation with the Troll Score.

Clustering based on visited state-action pairs   We perform user clustering considering the visited state-action pairs as a set of features, which, therefore, consists of the 11 possible combinations of states and actions. For every account, the values of the features are either set to 1, if the account visited a state-action pair, or to 0 otherwise. We use the Principal Component Analysis (PCA) to perform a dimensionality reduction, and we observe whether distinct clusters naturally emerge. In Fig. 5, we show the results of a PCA with two components, and we observe three clusters populated by both user and troll accounts. Figure 5(a) shows that users divide into three distinct clusters with slightly different distributions. Similarly, Fig. 5(b) shows three clusters for trolls, in which one cluster embeds the majority of troll accounts. For a better comparison, we display in Fig. 5(c) the joint plot of users and trolls combined, from which we can appreciate the three distinct *behavioral clusters*, where trolls and users coexist.

Table 2 reports the number of trolls and users present in each of the three clusters. Cluster 1 refers to the largest cluster (bottom left in Fig. 5), Cluster 2 refers to the cluster on the bottom right, and Cluster 3 refers to the one on the top right. By observing the visited state-action pairs, the three clusters can be differentiated as follows:

- Cluster 1: Accounts in this cluster tweet or stay silent with any received feedback;
- Cluster 2: Accounts in this cluster tweet when no feedback is provided by the environment (state NT) or they retweet with any received feedback;
- Cluster 3: Accounts in this cluster tweet or interact with others (replying to or mentioning other accounts) with any received feedback.

Our findings are consistent with [49] and show that most trolls (trolls in Clusters 1 and 2) tweet and retweet regardless of the received feedback, while a very small percentage of them (Cluster 3) participate in discussions by means of replies and mentions. This behavior is significantly different with respect to that of user accounts, as they are more evenly distributed among the three clusters and tend to join and participate more in discussions.
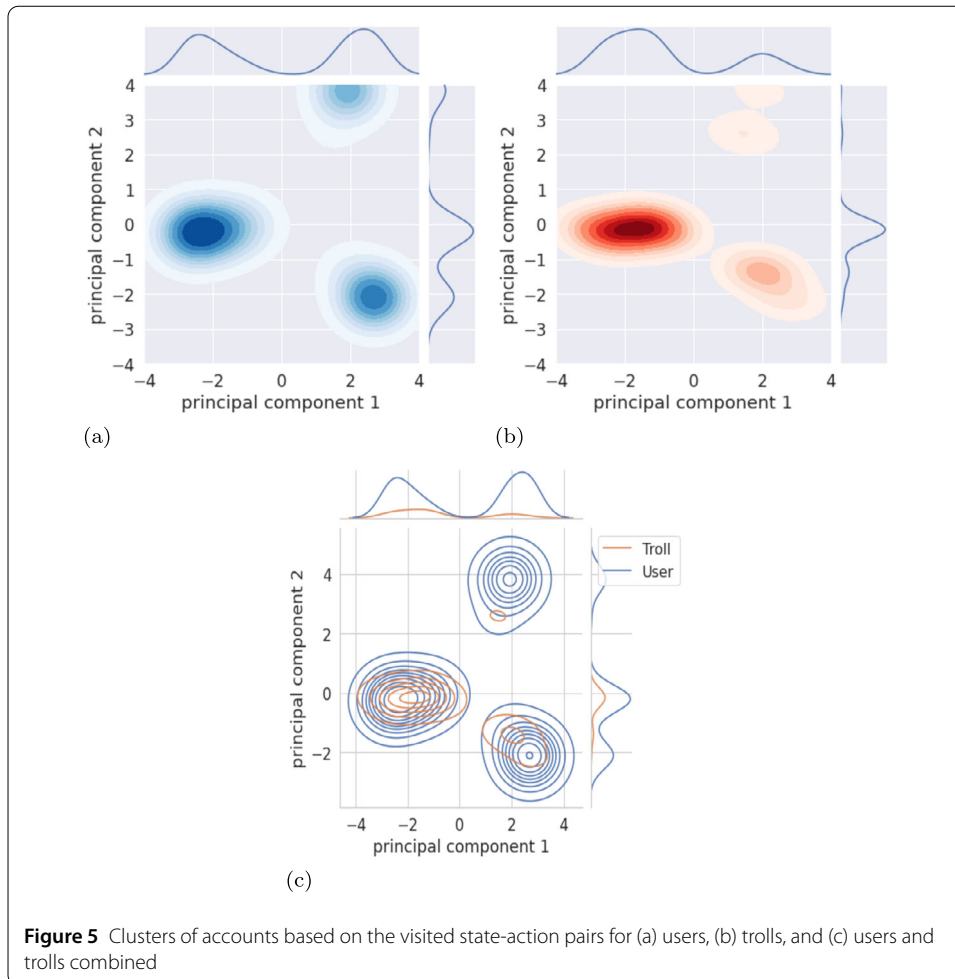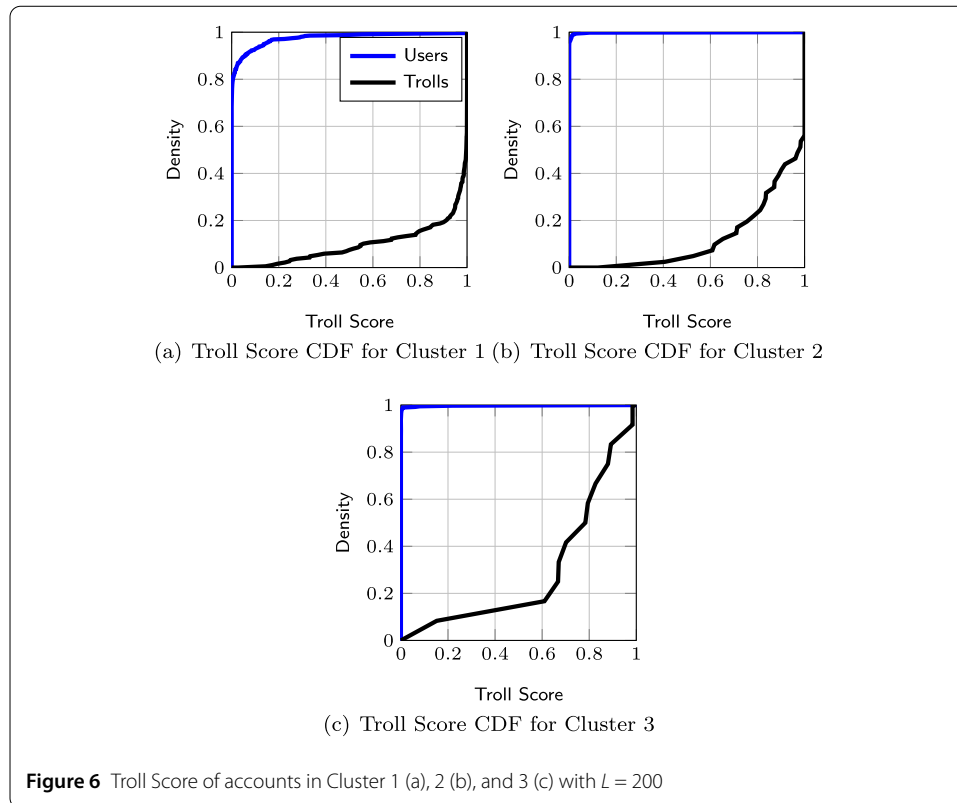
Figure 5 Clusters of accounts based on the visited state-action pairs for (a) users, (b) trolls, and (c) users and trolls combined

Table 2 Distribution of user and troll accounts in the three naturally-emerging clusters

	Cluster 1	Cluster 2	Cluster 3
Users	977 (49.5%)	603 (30.43%)	401 (20.24%)
Trolls	256 (75.07%)	65 (19.06%)	20 (5.86%)

Troll score per cluster   To further inspect the behavior of users and trolls in the three clusters, we evaluate the Troll Score of the accounts belonging to each of these clusters separately. Figures 6(a)-(c) show the CDF of the Troll Score of accounts in Clusters 1-3, respectively. Figures show that the CDF of the Troll Score of users in all clusters has an *exponential* shape, indicating that most users have a near-zero Troll Score. In contrast, the CDF of the Troll Score of trolls in every cluster shows a *logarithmic* shape, suggesting that most of the trolls have a high Troll Score. However, the CDF of trolls within every cluster has notable differences.

For instance, in Clusters 1 and 2, most trolls are characterized by high Troll Scores (e.g., only about 20% of trolls have a Troll Score below 0.8), while in Cluster 3, 60% of trolls have a Troll Score below 0.8. This suggests that trolls belonging to Cluster 3 have a relatively low Troll Score, indicating that a relevant part of their trajectories is classified as user trajectories. This is likely due to the peculiar behavior of the small set of troll accounts

(a) Troll Score CDF for Cluster 1 (b) Troll Score CDF for Cluster 2

(c) Troll Score CDF for Cluster 3

Figure 6 Troll Score of accounts in Cluster 1 (a), 2 (b), and 3 (c) with $L = 200$

(5%) belonging to Cluster 3. Based on this observation, we argue that our LSTM model identifies specific patterns of activities of trolls in Clusters 1 and 2, as they represent the most populated clusters (95% of the accounts).

6.2.3 Troll score-based classification vs. existing approaches

In this Section, we evaluate the classification performance of our method against other existing approaches. In our proposed methodology, the training set is used to find the *optimal* Troll Score threshold, while the test set is used to evaluate the classification. The value of L is set to 200, based on the results in Sect. 6.1.3. We evaluate the performance of our proposed Troll Score-based approach by comparing it to (*i*) a behavioral approach based on Inverse Reinforcement Learning (IRL) [49], which is the only other language-agnostic solution that relies solely on the sequences of accounts' sharing activity to detect trolls.; and (*ii*) two linguistic approaches [40, 45] that use the text of the shared content to extract linguistic features for identifying troll accounts.

Table 3 displays the AUC, recall, precision, and F1-score of the behavioral and linguistic approaches tested through a stratified 10-fold cross-validation. Two facts are worth noting. First, the Troll Score-based solution outperforms the other behavioral approach [49] with an AUC of 90.6%, a precision of 90.1%, and a recall of 89.6%. Second, the linguistic approaches achieve high classification accuracy, with all the metrics close to 100%. As suggested by prior research [65, 66], the high classification accuracy of linguistic approaches is likely due to patterns of English misuse of content and function words among non-native English speakers, such as Russian trolls, which allows for identifying their original spoken language even when the speaker is fluent [40]. However, it is noteworthy that our Troll

Table 3 Comparison between behavioral and linguistic approaches for troll account identification

Metric	Behavioral approaches		Linguistic approaches	
	Troll score-based	IRL-based [49]	Addawood et al. [45]	Im et al. [40]
Accuracy	**90.6**	83.0	98.9	97.7
AUC	**90.5**	89.1	99.8	99.7
Precision	**90.1**	84.0	98.5	97.5
Recall	**89.6**	85.0	97.3	97.5
F1-Score	**89.7**	84.5	98.9	97.5

Score-based approach, which does not have access to the textual content of shared messages, approaches the performance of linguistic approaches that leverage message content. This represents a valuable advantage of behavioral models, especially considering the increasing capabilities of LLMs to refine and correct human-generated text. In the future, it will be crucial to develop models that use behavioral cues to detect deceptive activities, as distinguishing between content generated or revised by LLMs and human actors is becoming increasingly challenging [50, 51]. Based on these premises, in the next Section, we propose to investigate the generalizability of our approach to identifying drivers of information operations,[3] which might employ a more sophisticated combination of human curation, automation through social bots, and LLM-powered content generation [29, 47].

6.3 Identification of drivers of information operations

In this Section, we investigate whether the proposed approach can effectively identify a diverse set of actors involved in influence campaigns. Similarly to Nwala et al. [29], we refer to these malicious entities as "drivers of information operations" (*IO drivers* for short). Specifically, we evaluate our approach's effectiveness in detecting IO drivers involved in discussions related to the 2020 U.S. Presidential election on Twitter. To conduct our analysis, we use the dataset made available by Chen et al. [67], which includes tweets that mention specific election-related keywords. We identify IO drivers in this dataset by consulting the list of accounts involved in state-linked information operations released by Twitter [27].

The rationale behind interleaving these two data sources is twofold. First, we are interested in observing orchestrated interference campaigns originating from different countries, similar to Russian meddling in the 2016 U.S. election. As shown in Fig. 7, we uncovered influence efforts from a large set of information operations, with the majority of IO drivers operating in campaigns linked to Iran, Cuba, and Uganda. Notably, although these accounts originate from different countries, they almost exclusively share content in English, with 90% of their tweets being in English. Second, Twitter IO archive [27] includes only messages shared by IO drivers during their lifespan on Twitter, which requires collecting another dataset to produce the negative class. This, in turn, may introduce biases in the data collection, primarily due to the selection of organic users and their potentially diverse activity on Twitter. Additionally, Twitter data only incorporates IO drivers' *active online activities*, whereas our methodology leverages the feedback received by these accounts (*passive online activities*). However, as these accounts have been suspended from Twitter, it is now impossible to collect organic users' interactions with IO drivers. Overall, the proposed study case represents a more challenging scenario with respect to the
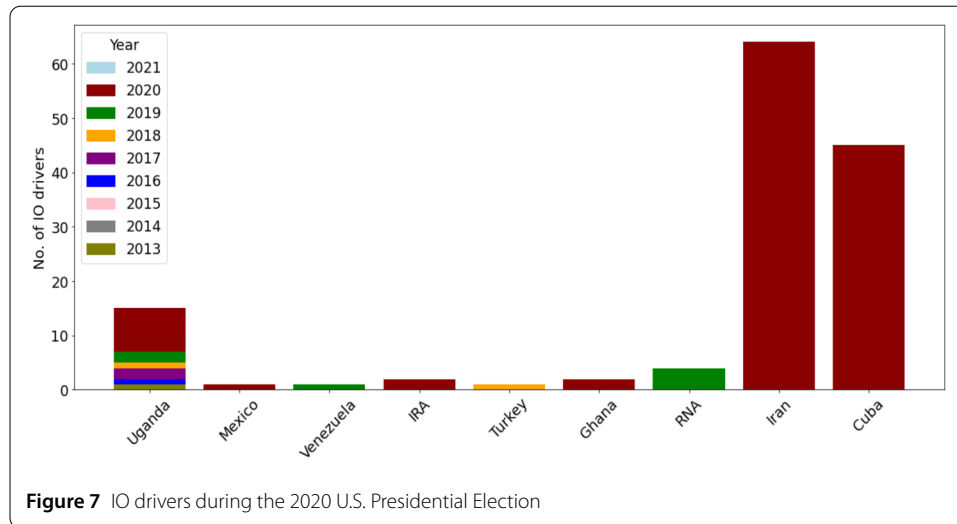
---

[3]https://help.twitter.com/en/rules-and-policies/platform-manipulation.

Figure 7 IO drivers during the 2020 U.S. Presidential Election

Table 4 Comparison between behavioral and linguistic approaches for the identification of IO drivers

Metric	Behavioral approaches		Linguistic approaches	
	Troll score-based	IRL-based [49]	Addawood et al. [45]	Im et al. [40]
Accuracy	**79.5**	75.6	93.4	91.5
AUC	**80.1**	81.8	98.3	96.9
Precision	**83.4**	76.9	92.0	90.8
Recall	**81.4**	73.4	90.0	88.1
F1-Score	**81.6**	74.8	90.9	89.2

detection of IO drivers using Twitter data, as we do not benefit from the whole history of IO drivers' sharing activity on Twitter, but we can only observe their actions and received interactions during the 2020 U.S. election.

Similarly to the previous study case, we select accounts involved in at least ten active and passive online activities. We find 145 IO driver accounts that we aim to classify against a set of 400 randomly selected organic users. We generated trajectories considering $L = 100$, which lead to a total of 2479 organic users' trajectories and 900 IO drivers' trajectories. By following our approach, we perform a classification of the two classes of accounts, whose results are reported in Table 4. Our approach achieves promising performance, with an AUC of 81.8%, confirming its capability and potential to generalize to different drivers of influence campaigns operating in various countries. It is worth noting that, even in this challenging scenario, our proposed methodology surpasses the other behavioral approach [49] in terms of precision, recall, and F1-Score. While it does not achieve the same classification performance as linguistic approaches, this outcome is promising and underscores the generalization capability of our approach, all while maintaining the characteristic of not requiring access to any content shared in users' messages.

Finally, Fig. 8 displays the Troll Score of the analyzed users and IO drivers. It can be observed that our approach clearly identifies about 70% of users and IO driver accounts that have a Troll Score of 0 and 1, respectively. However, the remaining fractions of IO drivers and users are more difficult to distinguish, leading to a natural trade-off between true and false positive rates. It is worth noting that our threshold can be adjusted to build a more conservative model and minimize inaccurate classifications of organic users. This

Figure 8 Troll Score CDF of users and IO drivers with $L = 100$
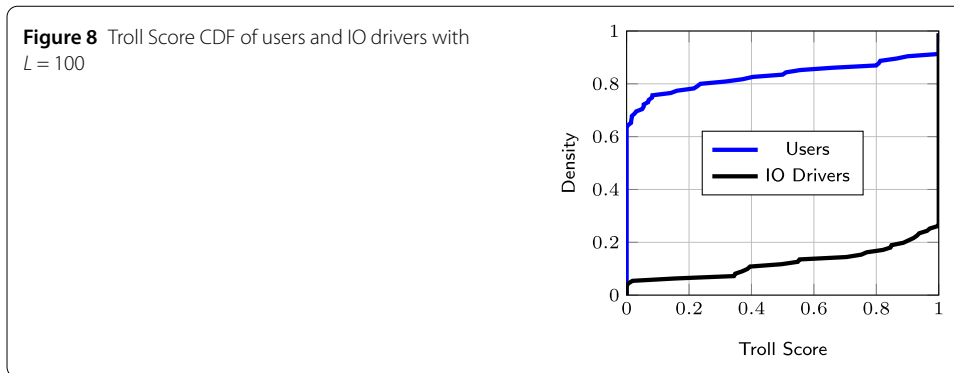
is particularly important to reduce social media providers' overhead in the case of misclassification, which may result in moderation intervention such as suspension or ban. Moreover, the different patterns in Fig. 8 compared to Fig. 4 may reflect a behavioral evolution in IO drivers' deceptive actions, including coordinated strategies and automation [15, 47, 48].

7 Conclusion

In this paper, we introduced a novel two-step AI approach for the detection of troll accounts based solely on behavioral cues. The first step employs an LSTM neural network to classify sequences of online activities as either troll or user activity. In the second step, we utilize the classified sequences to calculate a metric named the "Troll Score", which quantifies the extent to which an account exhibits troll-like behavior. Our initial experiments primarily focused on identifying Russian troll activity on Twitter during the 2016 U.S. election. The results show that our approach identifies account sequences with an AUC of nearly 99% and, accordingly, classifies troll and user accounts with an AUC of 91%, outperforming existing behavioral approaches. Although it does not reach the same level of performance as language-based techniques, our proposed approach does not require access to any content shared in users' messages, making it particularly robust in the era of growing LLM usage for influence campaigns. When analyzing the Troll Score of the examined accounts, we observed that the majority of trolls exhibit a high Troll Score, indicating that most of their trajectories are classified as troll activity, while most users display a low Troll Score, suggesting that their trajectories are predominantly classified as user activity. However, intriguingly, the results also reveal that a small fraction of trolls have a low Troll Score, implying that their activity closely resembles the behavior of organic users. To delve deeper into this finding, we conducted a detailed analysis of trolls and users based on their visited state-action pairs. Our analysis reveals the presence of three distinct behavioral clusters among the examined accounts, each populated by both trolls and users. Interestingly, trolls within a specific, albeit less populated, cluster exhibit a relatively lower Troll Score compared to trolls in other clusters, indicating that this particular group of trolls behaves more similarly to organic users than other trolls.

Finally, we tested our approach in a different context with the aim of assessing its generalizability to other scenarios and deceptive, coordinated activities. Specifically, we evaluated the performance of our methodology in identifying accounts involved in diverse information operations during the 2020 U.S. election discussion on Twitter. The results reveal that, while the nature of the drivers of these campaigns might vary, including bots and

automation, our methodology effectively identifies their activity and produces promising classification results, which will be further validated in our future endeavors.

Abbreviations
SMNs, Social Media Networks; IRA, Internet Research Agency; LSTM, Long short-term memory; IRL, Inverse Reinforcement Learning; MDP, Markov Decision Process; SVM, Support Vector Machine; ANN, Artificial Neural Network; KNN, K-Nearest Neighbors; CDF, Cumulative Distribution Function; PCA, Principle Component Analysis; IO, Information Operation.

Availability of data and materials
The code and scripts used to implement the methodological framework detailed below are freely available to the research community. Github Repository:
https://github.com/FatimaEzzedinee/Exposing-Influence-Campaigns-in-the-Age-of-LLMs-A-Behavioral-Based-AI-Approach. In compliance with Twitter's terms of service, we will grant complete access to the tweet IDs that were examined in our study, which can be retrieved using Twitter's API. It should be noted, however, that tweets from trolls may be inaccessible due to the removal of their accounts and associated tweets from Twitter.

Declarations

Competing interests
The authors declare that they have no competing interests.

Author contributions
FE created the software of this study, conceptualized the LSTM-architecture for classification, and drafted the manuscript. OA conceived the framework, carried out the analysis, and drafted the manuscript. LL designed and supervised the study, conceived the framework and its implementation, and finalized the manuscript across the review rounds. SG designed the study, analyzed the results, and finalized the manuscript. GN contributed to the software implementation, gathered the dataset, and drafted the manuscript. EF participated in the analysis of the results and finalized the manuscript. IS participated in the design of the study and in the writing of the manuscript. All authors read and approved the final manuscript.

Author details
[1]Department of Innovative Technologies, University of Applied Sciences and Arts of Southern Switzerland, Lugano, Switzerland. [2]Department of Applied Mathematics, Faculty of Science, Lebanese University, Beirut, Lebanon. [3]Information Sciences Institute, Viterbi School of Engineering, University of Southern California, Marina del Rey, CA, USA.

References
1. Luceri L, Cresci S, Giordano S (2021) Social media against society. The Internet and the 2020 Campaign, 1
2. Aro J (2016) The cyberspace war: propaganda and trolling as warfare tools. Eur View 15(1):121–132
3. Zollo F, Novak PK, Del Vicario M, Bessi A, Mozetič I, Scala A, Caldarelli G, Quattrociocchi W (2015) Emotional dynamics in the age of misinformation. PLoS ONE 10(9):0138740
4. Pariser E (2011) The filter bubble: what the Internet is hiding from you. Penguin, New York
5. Luceri L, Cardoso F, Giordano S (2021) Down the bot hole: actionable insights from a one-year analysis of bot activity on Twitter. First Monday
6. Pierri F, Perry BL, DeVerna MR, Yang K-C, Flammini A, Menczer F, Bryden J (2022) Online misinformation is linked to early COVID-19 vaccination hesitancy and refusal. Sci Rep 12(1):1–7
7. Ferrara E (2015) "Manipulation and abuse on social media" by Emilio Ferrara with Ching-Man Au Yeung as coordinator. ACM SIGWEB Newsl 2015:4
8. Pierri F, Luceri L, Jindal N, Ferrara E (2023) Propaganda and misinformation on Facebook and Twitter during the Russian invasion of Ukraine. In: Proceedings of the 15th ACM web science conference 2023, pp 65–74
9. Ferrara E, Cresci S, Luceri L (2020) Misinformation, manipulation, and abuse on social media in the era of COVID-19. J Comput Soc Sci 3(2):271–277
10. Diseases TLI (2020) The COVID-19 infodemic. Lancet Infect Dis 20(8):875
11. Hu Z, Yang Z, Li Q, Zhang A (2020) The COVID-19 infodemic: infodemiology study analyzing stigmatizing search terms. J Med Internet Res 22(11):22639
12. Nogara G, Vishnuprasad PS, Cardoso F, Ayoub O, Giordano S, Luceri L (2022) The disinformation dozen: an exploratory analysis of COVID-19 disinformation proliferation on Twitter. In: 14th ACM web science conference 2022, pp 348–358
13. Pierri F, DeVerna MR, Yang K-C, Axelrod D, Bryden J, Menczer F (2022) One year of COVID-19 vaccine misinformation on Twitter. arXiv preprint. arXiv:2209.01675
14. Wang EL, Luceri L, Pierri F, Ferrara E (2022) Identifying and characterizing behavioral classes of radicalization within the qanon conspiracy on Twitter. arXiv preprint. arXiv:2209.09339
15. Suresh VP, Nogara G, Cardoso F, Cresci S, Giordano S, Luceri L (2024) Tracking fringe and coordinated activity on Twitter leading up to the us capitol attack. In: Proceedings of the international AAAI conference on web and social media

16. Phadke S, Samory M, Mitra T (2022) Pathways through conspiracy: the evolution of conspiracy radicalization through engagement in online conspiracy discussions. In: Proceedings of the international AAAI conference on web and social media, vol 16, pp 770–781

17. Allem J-P, Ferrara E, Uppu SP, Cruz TB, Unger JB (2017) E-cigarette surveillance with social media data: social bots, emerging topics, and trends. JMIR Public Health Surveill 3(4):8641

18. Del Vicario M, Vivaldo G, Bessi A, Zollo F, Scala A, Caldarelli G, Quattrociocchi W (2016) Echo chambers: emotional contagion and group polarization on Facebook. Sci Rep 6(1):1–12

19. Matakos A, Terzi E, Tsaparas P (2017) Measuring and moderating opinion polarization in social networks. Data Min Knowl Discov 31(5):1480–1505

20. Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. Science 359(6380):1146–1151

21. Metaxas PT, Mustafaraj E (2012) Social media and the elections. Science 338(6106):472–473

22. Gatta VL, Luceri L, Fabbri F, Ferrara E (2023) The interconnected nature of online harm and moderation: investigating the cross-platform spread of harmful content between youtube and Twitter. In: Proceedings of the 34th ACM conference on hypertext and social media, pp 1–10

23. Carroll O (2017) St. petersburg troll farm had 90 dedicated staff working to influence US election campaign. The Independent

24. Popken B (2018) Twitter deleted Russian troll tweets. So we published more than 200,000 of them. NBC News 14

25. Mueller RS (2019) The Mueller report: report on the investigation into Russian interference in the 2016 presidential election. WSBLD

26. Lopez J, Hillygus DS (2018) Why so serious?: survey trolls and misinformation. Why so serious

27. Gadde V, Beykpour K (2020) Additional steps we're taking ahead of the 2020 US election. Social Media Twitter

28. Alizadeh M, Shapiro JN, Buntain C, Tucker JA (2020) Content-based features predict social media influence operations. Sci Adv 6(30):5824

29. Nwala AC, Flammini A, Menczer F (2023) A language framework for modeling social media account behavior. EPJ Data Sci 12(1):33

30. Pierri F, Luceri L, Ferrara E (2022) How does Twitter account moderation work? Dynamics of account creation and suspension during major geopolitical events. arXiv preprint. arXiv:2209.07614

31. Luceri L, Deb A, Giordano S, Ferrara E (2019) Evolution of bot and human behavior during elections. First Monday

32. Kudugunta S, Ferrara E (2018) Deep neural networks for bot detection. Inf Sci 467:312–322

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

34. 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

35. Chavoshi N, Hamooni H, Mueen A (2016) Debot: Twitter bot detection via warped correlation. In: Icdm, pp 817–822

36. Abou Daya A, Salahuddin MA, Limam N, Boutaba R (2019) A graph-based machine learning approach for bot detection. In: 2019 IFIP/IEEE symposium on integrated network and service management (IM). IEEE, New York, pp 144–152

37. Cresci S, Di Pietro R, Petrocchi M, Spognardi A, Tesconi M (2017) Social fingerprinting: detection of spambot groups through DNA-inspired behavioral modeling. IEEE Trans Dependable Secure Comput 15(4):561–576

38. Ferrara E (2022) Twitter spam and false accounts prevalence, detection and characterization: a survey. arXiv preprint. arXiv:2211.05913

39. Zannettou S, Caulfield T, Setzer W, Sirivianos M, Stringhini G, Blackburn J (2019) Who let the trolls out? Towards understanding state-sponsored trolls. In: Proceedings of the 10th ACM conference on web science, pp 353–362

40. Im J, Chandrasekharan E, Sargent J, Lighthammer P, Denby T, Bhargava A, Hemphill L, Jurgens D, Gilbert E (2020) Still out there: modeling and identifying Russian troll accounts on Twitter. In: 12th ACM conference on web science, pp 1–10

41. Badawy A, Addawood A, Lerman K, Ferrara E (2019) Characterizing the 2016 Russian IRA influence campaign. Soc Netw Anal Min 9(1):1–11

42. Alhazbi S (2020) Behavior-based machine learning approaches to identify state-sponsored trolls on Twitter. IEEE Access 8:195132–195141

43. Saeed MH, Ali S, Blackburn J, De Cristofaro E, Zannettou S, Stringhini G (2021) Trollmagnifier: detecting state-sponsored troll accounts on Reddit. arXiv preprint. arXiv:2112.00443

44. Mazza M, Avvenuti M, Cresci S, Tesconi M (2022) Investigating the difference between trolls, social bots, and humans on Twitter. Comput Commun 196:23–36

45. Addawood A, Badawy A, Lerman K, Ferrara E (2019) Linguistic cues to deception: identifying political trolls on social media. In: Proceedings of the international AAAI conference on web and social media, vol 13, pp 15–25

46. Jachim P, Sharevski F, Treebridge P (2020) Trollhunter [evader]: automated detection [evasion] of Twitter trolls during the COVID-19 pandemic. In: New security paradigms workshop 2020, pp 59–75

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

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

49. Luceri L, Giordano S, Ferrara E (2020) 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, vol 14, pp 417–427

50. Menczer F, Crandall D, Ahn Y-Y, Kapadia A (2023) Addressing the harms of AI-generated inauthentic content. Nat Mach Intell 5:679–680

51. Mitrović S, Andreoletti D, Ayoub O (2023) Chatgpt or human? Detect and explain. Explaining decisions of machine learning model for detecting short ChatGPT-generated text. arXiv preprint. arXiv:2301.13852

52. Frommer D (2019) Twitter's list of 2,752 Russian trolls

53. Weller H, Woo J (2019) Identifying Russian trolls on Reddit with deep learning and bert word embeddings

54. Vanhove T, Leroux P, Wauters T, De Turck F (2013) Towards the design of a platform for abuse detection in OSNs using multimedial data analysis. In: 2013 IFIP/IEEE international symposium on integrated network management (IM 2013). IEEE, New York, pp 1195–1198

55. Valldor E, Stenborg K, Gustavsson D (2018) Firearm detection in social media images. In: Swedish symposium on deep learning

56. Kim D, Graham T, Wan Z, Rizoiu M-A (2019) 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):331–351
57. Wang G, Zhang X, Tang S, Zheng H, Zhao BY (2016) Unsupervised clickstream clustering for user behavior analysis. In: Proceedings of the 2016 CHI conference on human factors in computing systems, pp 225–236
58. Metaxas P, Mustafaraj E, Wong K, Zeng L, O'Keefe M, Finn S (2015) What do retweets indicate? Results from user survey and meta-review of research. In: Proceedings of the international AAAI conference on web and social media, vol 9
59. Stella M, Ferrara E, De Domenico M (2018) Bots increase exposure to negative and inflammatory content in online social systems. Proc Natl Acad Sci 115(49):12435–12440
60. Bessi A, Ferrara E (2016) Social bots distort the 2016 US presidential election online discussion. First monday 21(11-7)
61. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15(1):1929–1958
62. Nielsen MA (2015) Neural networks and deep learning
63. Liu X-Y, Wu J, Zhou Z-H (2008) Exploratory undersampling for class-imbalance learning. IEEE Trans Syst Man Cybern, Part B, Cybern 39(2):539–550
64. Drumnond C (2003) Class imbalance and cost sensitivity: why undersampling beats oversampling. In: ICML-KDD 2003 workshop: learning from imbalanced datasets, vol 3
65. Ionin T, Zubizarreta ML, Maldonado SB (2008) Sources of linguistic knowledge in the second language acquisition of English articles. Lingua 118(4):554–576
66. Nicolai G, Kondrak G (2014) Does the phonology of l1 show up in l2 texts? In: Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: short papers), pp 854–859
67. Chen E, Deb A, Ferrara E (2021) #Election2020: the first public Twitter dataset on the 2020 US presidential election. J Comput Soc Sci 5(1):1–18

Publisher's Note

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