


Received October 4, 2020, accepted October 18, 2020, date of publication October 26, 2020, date of current version November 6, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3033666

Behavior-Based Machine Learning Approaches to Identify State-Sponsored Trolls on Twitter

SALEH ALHAZBI , (Senior Member, IEEE)

Department of Computer Science and Engineering, College of Engineering, Qatar University, Doha, Qatar

e-mail: salhazbi@qu.edu.qa

Open Access funding provided by the Qatar National Library.

ABSTRACT In recent years, there has been an increased prevalence of adopting state-sponsored trolls by governments and political organizations to influence public opinion through disinformation campaigns on social media platforms. This phenomenon negatively affects the political process, causes distrust in the political systems, sows discord within societies, and hastens political polarization. Thus, there is a need to develop automated approaches to identify sponsored-troll accounts on social media in order to mitigate their impacts on the political process and to protect people against opinion manipulation. In this paper, we argue that behaviors of sponsored-troll accounts on social media are different from ordinary users' because of their extrinsic motivation, and they cannot completely hide their suspicious behaviors, therefore these accounts can be identified using machine learning approaches based solely on their behaviors on the social media platforms. We have proposed a set of behavioral features of users' activities on Twitter. Based on these features, we developed four classification models to identify political troll accounts, these models are based on decision tree, random forest, Adaboost, and gradient boost algorithms. The models were trained and evaluated on a set of Saudi trolls disclosed by Twitter in 2019, the overall classification accuracy reaches up to 94.4%. The models also are capable to identify the Russian trolls with accuracy up to 72.6% without training on this set of trolls. This indicates that although the strategies of coordinated trolls might vary from an organization to another, they are all just employees and have common behaviors that can be identified.

INDEX TERMS State-sponsored trolls, disinformation, propaganda, behavioral pattern.


I. INTRODUCTION

Social networks increasingly have become a vital tool for disseminating opinions and real-time information, which promotes democracy and increase the empowerment of citizens. They give everyone a voice in governments to discuss public issues, organize social movements, and hold leaders accountable [1], [2]. On the other hand, there has been a growing concern about using social networks as tools for social control and public opinion manipulation.

Many governments now devote significant resources for social network manipulation using fake accounts in order to spread computational propaganda that supports their agenda [3], these accounts can be a bot, cyborg, or human. Social bots are social network accounts that are controlled completely by a computer program to mimic human behavior online. They can be created in enormous numbers and used to share/retweet information such as news [4], influence public opinion, amplify contents, or drown out political

dissent [5]. Unlike bot and in the middle between humans and bots, cyborgs are fake accounts that combine automation and human curation, after an account is created by human, automated programs might be used to post and share information, at the same time user might participate to tweet and interact with other users. Cyborgs can be defined as bot-assisted humans or human-assisted bots [6], [7]. Recently, authoritarian regimes employ significant numbers of people to coordinately use social networks to manipulate public opinion by targeting local audiences or foreign publics [5], this new phenomenon is known as state-sponsored trolling.

Generally, the term "trolling" is used widely to refer to different types of online disruptive activities, it classically refers to people who post online inflammatory remarks in order to provoke the desired reaction [8]. However, this definition related the motivation of this phenomenon to the personal psychological needs of troll, which is different from state-sponsored trolls whose main objective is to propagate political agenda [9]. Moreover, they are different in terms of behavioral factors, where state-sponsored trolls intensively reposted messages from different accounts or nicknames and

The associate editor coordinating the review of this manuscript and approving it for publication was Feng Xia .

republish information and links in order to strongly support a particular political stance [10]. Multiple terms are used to refer to pro-government online trolls like “cyber troops” [5], “cyber army” [11], “troll farm” [12], and “troll factories” [13].

A. STATE-SPONSORED TROLLS

The Russian interference in the 2016 US presidential election revealed another dangerous type of cyberwarfare, which leverages social networks to propagate disinformation and manipulate public opinion. According to Mueller reports [14], many accounts on Twitter and Facebook were linked to an organization known as the Internet Research Agency (IRA), a Russian company that creates online propaganda, formed a well-designed campaign to influence political public attitudes and behaviors in the US. In 2017, Facebook identified 470 IRA-controlled accounts that made around 80,000 posts, which reached around 126 million persons. In 2018, Twitter identified 3814 accounts linked to IRA and indicated that around 1.4 million users engaged with these accounts’ tweets leading to nearly 73 million engagements.

Many researchers have studied the activities of these accounts [12], [15], [16], it is found they were designed to promote political polarization, sow divisions among citizens, and spread fake news. Moreover, the influence of these accounts reached major American news outlets and formal politicians [17]. The IRA troll farm represents the first revealed case of hiring human operators to carry a deceptive online interference campaign. Lately, more countries and organizations use the same technique, according to [3], the number of countries that utilize social media to organize computational propaganda has been increased from 28 countries in 2017 to 48 in 2018 and 70 in 2019. These countries employ cyber troops through governmental agencies like communication and digital ministries or military-led campaign to shape local public opinion and discredit political opponents. Additionally, some countries use such cyber troops to engage in foreign influence operations.

In addition to the 2016 US presidential election, researchers discovered other political events that accompanied by computational propaganda managed by social sponsored-trolls, these include the 2016 UK Brexit referendum [18], the 2017 French presidential election [19], and the Gulf crisis [20]. Generally, state-sponsored trolling has become a global phenomenon over the last four years. Because of their ability to influence public opinions, governments also dedicated trolls to bring attention to their agenda and to target local individuals or organizations that criticize the government [5].

Overall, state-sponsored trolling represents a toxic for democracy, it negatively affects the political process, causes distrust in the political systems [16], sows discord within societies, hastens political polarization [17], and influence electoral outcomes. Accordingly, there is a need to develop automated approaches to identify sponsored-troll accounts in

order to mitigate their impacts on the political process and to protect people against opinion manipulation.

In this paper, we aim to use machine-learning algorithms to identify state-sponsored accounts on Twitter based solely on their behavioral features extracted from 500 consecutive tweets. These features include the daily average number of tweets, using hashtags and URLs, retweets and replies, and temporal patterns. The main contributions of this paper as follows:

- Our work is based on the assumption that state-sponsored trolls are just employees, who hired and are paid by governments to post messages and spread propaganda that supports the employers’ agenda. Therefore, they are not intrinsically motivated, so their behaviors on social media will be different from ordinary users’ ones and can be detected using machine learning algorithms. Moreover and based on that assumption, these trolls will have common online behaviors on social networks regardless of their languages, or the organization they support, or the political issue they target. Hence, we can build general models that are capable to identify these trolls without the need to analyze the contents of their post, which requires using natural language processing that will vary from one language to another.
- Unlike the previous works in this area that are merely based on IRA dataset trolls, our experiment is based on a dataset contains Saudi trolls disclosed by Twitter in December 2019. The details of this dataset is presented in section three.
- Although the proposed machine learning approaches have been trained and tested on the Saudi trolls set, they also have good performance when they are evaluated on the IRA trolls set. This indicates that sponsored-trolls have common behaviors regardless of the topic, or the language they used to post.
- As an application, our approach requires only the last 500 consecutive tweets to extract user’s behavior to classify the account if it is a troll or not. Extracting these tweets through Twitter API will not take much time, so classification will be in real-time.

The rest of this paper is organized as follows. Section II discusses related works in terms of users’ behaviors on Twitter and previous works to detect sponsored political trolls. Section III describes the method used in this work including the dataset and machine learning models. Section IV presents the results. Conclusion and future work are placed in section V.

II. RELATED WORK

In this section, we present related works, we start by identifying the variety of behaviors and activities that user can perform on Twitter, and discussing the theoretical background of our work, it is based on the assumption that state-sponsored trolls’ behaviors on Twitter are different from ordinary users. Then, we present previous works that use machine learning to identify state-sponsored trolls.

A. BEHAVIORAL PATTERNS OF POLITICAL TROLLS ON TWITTER

Twitter has become the most popular social network for analyzing the behavior of users when interacting online because researchers can access huge data easily via a number of Application Programming Interfaces (API). People on Twitter perform different types of activities, they publish their own contents, spread others' contents through retweets, or discuss others' contents through reply and mentioning [21]. Retweet has been considered a form of endorsement especially in political discussions, while replying is another way to express interest in the tweet and to carry on a conversation, and mentioning indicates that users highlight a message of a user in an attempt to start a conversation [22]. Additionally, people can include hashtags in their tweets in order to mark them as related to specific themes or events and to help other people find posts that are relevant to their interests. Furthermore, users can add URLs to their posts to direct readers to other websites that include further information that supports the contents of the tweets.

Our work is based on the concept that the behavior of sponsored-trolls on social media is mainly driven by extrinsic motivations [23], this is completely different from the behavior of genuine users whose online political participation is driven by intrinsic motivations relating to self-efficacy and empowerment [24]. Although political trolls try to mimic genuine users, they exhibit suspicious patterns of behaviors because of the extrinsic motivation and because they are centrally coordinated.

The authors in [23] used principal-agent theory to explain why sponsored-trolls cannot completely hide their suspicious behaviors. This theory is usually applied in business in order to conceptualize the information asymmetry between the principal (business owner), and the agent who works in favor of the principle. The principal does not have full information about how the agent will behave. The problem occurs because the goals of the principal and the agent are different, or at least the agent has less interest in the outcome than the principal [25].

Because sponsored-trolls are centrally coordinated by a principal(s), they simultaneously participate in organized campaigns by posting tweets about the same topic and within the same period, they simply retweet each other message or co-tweet the same messages independently. Asking the trolls (agents) to camouflage their activities in terms of type and post's contents requires extra works and effort. The principal can easily count the number of posts; however, assessing their qualities is difficult. Therefore, agents will create the required number of tweets without too much effort to camouflage their activities especially that usually one agent controls many accounts [23]. As a result of that, sponsored-trolls cannot hide their behavioral pattern completely, so by selecting and extracting indicative features, we can build machine-learning classifiers that can identify these accounts based on behavior-related features.

B. IDENTIFYING STATE-SPONSORED TROLLS

Most of the previous works related to internet trolls focus on traditional trollin; it is defined by Bishop [26] as "posting of any content on the internet which is provocative or offensive". The main intention of these trolls is to cause annoyance and trigger or increase the conflict for the purposes of their own amusement [8]. This is different from the intention of state-sponsored trolls who are employed by governments or organizations and following the agenda of their employers, not their own objectives. Previous research on traditional trolls studied different aspects of that phenomenon including trolls personalities [26], motivations [27], in addition, to develop machine learning models to identify these type of accounts [28]–[30].

Despite the increased prevalence of adopting state-sponsored trolls by governments and political organizations, only a few research addressed this phenomenon. The works in [5], [20], [23], [31], [32] focused only on analyzing trolls behaviors and evaluating their influence on the public, without addressing the challenge of identifying these accounts. On the other hand, there are multiple approaches for detecting social bots, but they are unlikely to identify trolls [33], [34], this makes identifying political trolls automatically is still an open challenge [35]. To our knowledge only [34], [36], and [37] studied different features of trolls to build automatic identification approaches, all of them built their experiments solely on IRA dataset trolls revealed by Twitter.

The authors in [34] developed a machine-learning model to identify state-supported trolls using a set of textual features. They also used the IRA dataset that disclosed by Twitter, however they only focused on accounts that use English as the main language because the goal was to detect IRA accounts that mimic regular US users, so the final list contains 2023 accounts. Their control accounts include 94,643 accounts that sampled randomly from accounts that use the English language, and have a location within the US. Additionally, the selected accounts should post at least 5 tweets between the 1st of August and 3^{1st} of December, 2016. They applied Latent Dirichlet Allocation (LDA) to extract the themes. After testing several classifiers, they found that Logistic Regression showed the best performance with 5-fold cross-validation, the performance is improved when extracted themes coupled with other text features. Using all features, the model reached precision of 96% and F1 score equal to 94%.

In [36], they developed three classifiers using Logistic regression, Decision Tree, and Adaptive Boosted Decision Tree. The troll accounts include 2286 out 3841 of IRA accounts disclosed by Twitter, only the accounts that used English as their main language were selected. The control accounts were randomly sampled and composed of 171,291 US-located accounts. Five groups of features were used: profile features, behavioral features, stop word usage features, language distributed features, and bag of words (BOW) features. The behavioral and linguistic features were

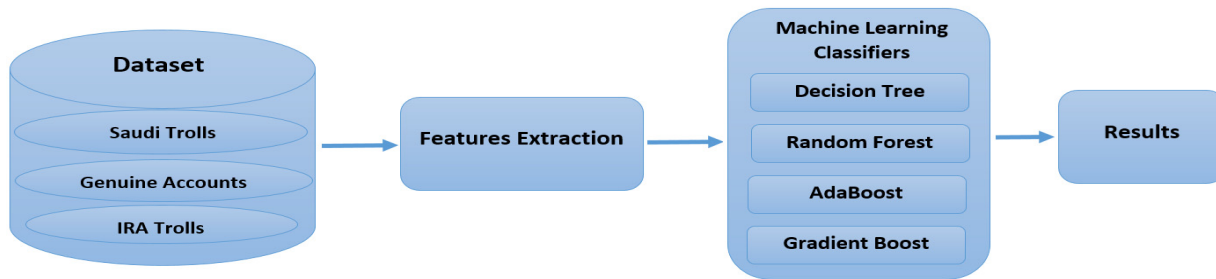


FIGURE 1. The methodology of this work.

extracted from the most recent 200 tweets of these accounts. With 10-fold cross-validation, the three classifiers had high accuracy (99%), however, the Adaptive Boosted Decision Tree perform the best precision (94%) comparing to Decision Tree (91%), and Logistic Regression(78%).

More recent work [37] employed Inverse Reinforcement Learning (IRL) to characterize the behavior of social accounts by inferring the reward structure behind their activities. This data are used as inputs to classification models to identify trolls. To build the dataset, the researchers relied initially on the IRA dataset. This data is filtered to include only users and trolls that shared at least 10 posts and were involved in at least 10 other accounts’ posts like retweet, reply, or mention. This yields 342 troll account and 1981 non-troll accounts. Different machine learning models were used. Adaboost achieved the best performance with an AUC of 89.1%.

III. METHODOLOGY

In this section, we describe our methodology including the dataset, feature extraction, and machine-learning algorithms we used in our work.

A. DATASE

To train and test machine-learning models to identify sponsored-trolls on Twitter, we need sponsored-trolls accounts and genuine (ordinary) accounts. Therefore, our dataset is composed of Saudi trolls, genuine accounts used as a control group to train and test the machine-learning classifiers. Additionally and in order to study our assumption that sponsored-trolls have common behaviors on Twitter irrespective of their organizations, or languages, we will test the proposed models with other types of trolls that have not been seen by the models during the training phase, so we use IRA trolls. Table 1 shows the number of accounts and tweets of the three types in our dataset.

TABLE 1. Description of the dataset.

ACCOUNT TYPE	ACCOUNTS	TWEETS
SAUDI TROLLS	1,681	840,500
GENUINE ACCOUNTS (CONTROL GROUP)	1,739	869,500
IRA TROLLS	752	376,000

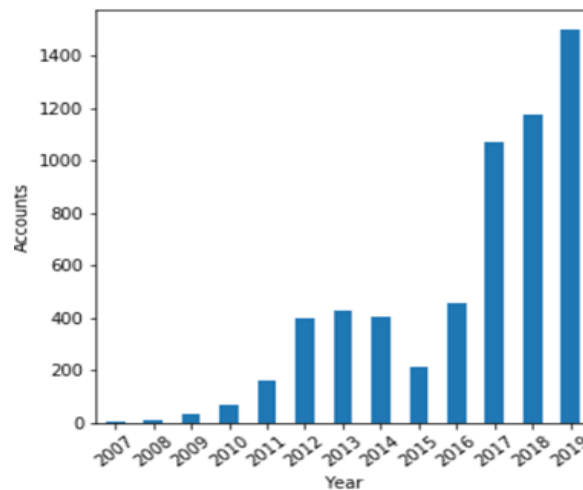


FIGURE 2. The Saudi trolls disclosed by Twitter.

1) THE SAUDI TROLL ACCOUNTS

We used Saudi suspended accounts, which were disclosed by Twitter in December 2019. This set includes 5929 accounts and over 32 million tweets. The accounts’ data include user id, user display name, user screen name, user location (based on user’s input), profile description, profile’s URL, number of followers, number of friends (following), date of account creation, and language of the account (based on user’s input).

The tweets’ data include tweet id, tweet language, tweet content, tweet time, if the tweet is a retweet or not, tweet’ id if the tweet is a replay to another tweet, URLs in the tweet, and hashtags in the tweet. Fig. 2 shows that around 63% of these accounts were created in 2017 and after that. This is related to the increased attention of social media impact in political propaganda. Moreover, it is related to the Gulf crisis that began in 2017, where frequently organized campaigns designed to support Saudi’s stance in this crisis [20].

In order to extract the pattern of tweeting behavior, we selected the accounts that have at least 500 tweets in the disclosed tweets set. This resulted in 1681 accounts. Fig. 3 shows the years when these accounts were created.

2) GENUINE ACCOUNTS

We manually compiled a list of personal genuine Saudi accounts on Twitter. To ensure that we only select genuine accounts, we targeted active personal accounts that are ver-

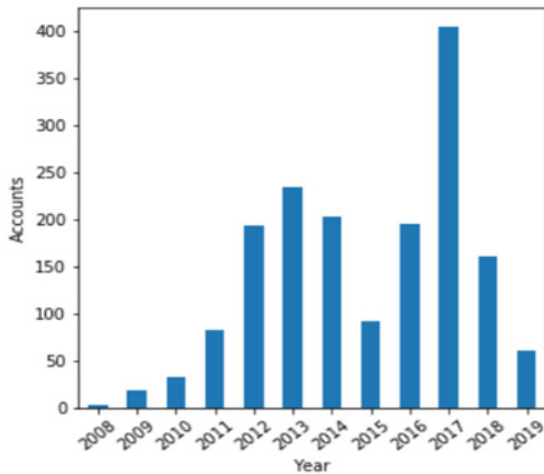


FIGURE 3. Saudi trolls selected in our dataset.

ified by Twitter, in addition to the accounts that belong to famous journalists, writers, politicians, or well-known professionals. All these users are from Saudi Arabia, and their language of tweeting is Arabic. We collected 2042 accounts, Fig. 4 illustrates the years of creating these accounts.

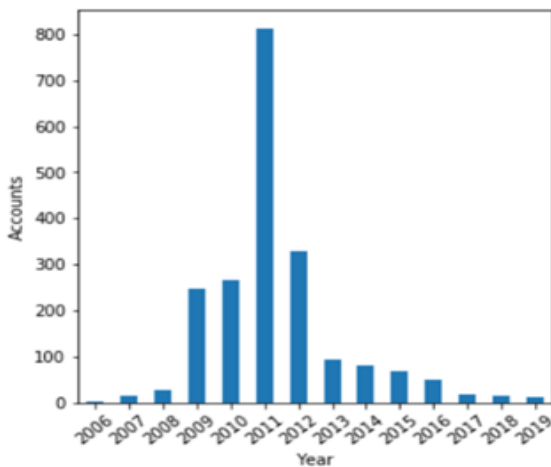


FIGURE 4. Collected genuine accounts.

We used “tweepy” package for Python as a tool for interacting with Twitter API to collect profiles data of the collected accounts as well as to extract the last 500 tweets for each account. We end up with 1739 accounts that have 500 tweets or more. The extracted data of each tweet includes user id, user display name, user screen name, user URL, user profile description, user account creation data, tweet text, tweet date and time, tweet language, URLs, and hashtags in the tweet, tweet’ id if the tweet is a replay to another tweet, and whether it is a retweet or not. Fig. 5 shows the years when these accounts were created.

3) THE RUSSIAN TROLLS ACCOUNTS

We use the suspended accounts disclosed by Twitter on Oct. 2018, these are the accounts that are controlled by a Russian government-linked agency called Internet Research Agency

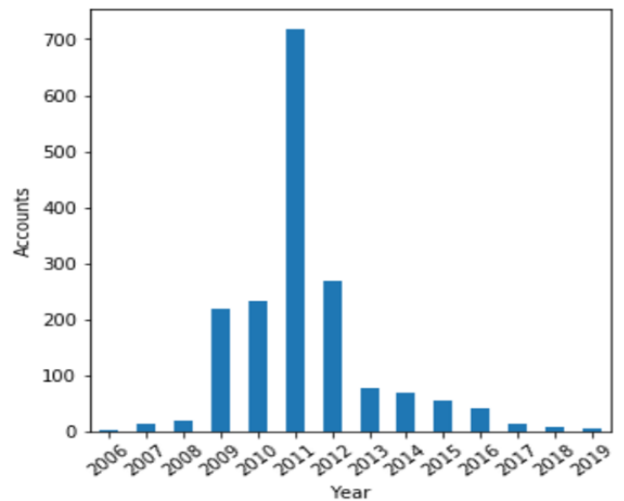


FIGURE 5. Genuine accounts in our dataset.

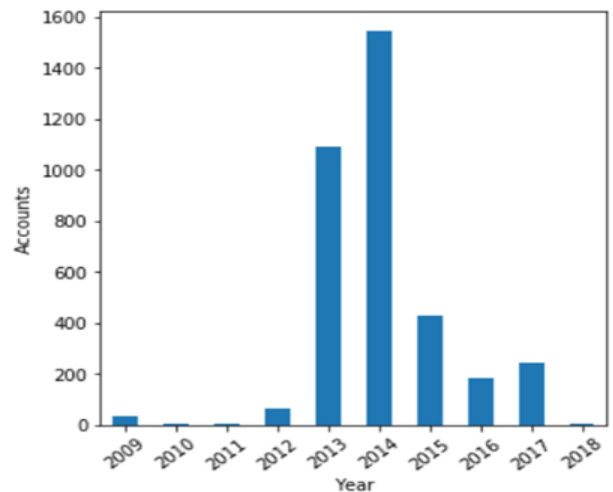


FIGURE 6. IRA trolls disclosed by Twitter.

(IRA). This set includes 3613 accounts and over nine million tweets. Fig. 6 shows the year when these accounts were created, most of these accounts created during 2013 and 2014.

We filter these accounts to select only the ones, which have at least 500 tweets, so we end up with 752 accounts. Fig 7 shows these accounts, 95% of these accounts were created between 2013 and 2015.

B. FEATURES ENGINEERING

Our objective in this work is to use users’ behaviors on Twitter as features for developing machine-learning models to classify state-supported trolls regardless of the contents of the tweets, we use the last 500 tweets to extract the features that represent user’ behavior. Our methodology is based on the following features:

1) AVERAGE NUMBER OF TWEETS PER DAY (AVG)

We use the date field of tweets to calculate the number of days of the period in which the user posts the 500 tweets, and calculate the average number of tweets per day.

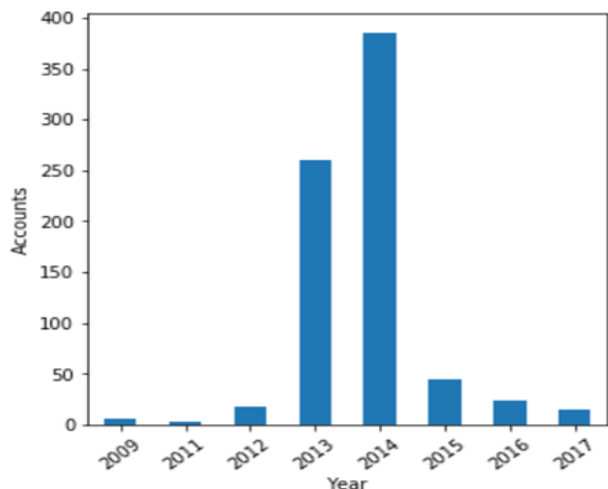


FIGURE 7. IRA accounts in our dataset.

2) STANDARD DEVIATION(STD)

In addition to the daily average number of tweets, we also calculate the standard deviation of the daily number of tweets.

3) NUMBER OF HASHTAGS (HST)

Users on Twitter use hashtags in order to create and follow a thread of discussion by using a word or a phrase preceded by hash character '#'. Hashtags are used by political trolls extensively to disseminate news, propaganda, and to flame up controversial topics like #BlackLivesMatter. It is found that the Russian trolls used hashtags in 32% of their tweets [31].

Trolls also use hashtags through well-organized campaigns to promote trends on Twitter that are aligned with the supporting government's agenda [13], [36], [38]. Therefore, the number of hashtags can be a distinctive feature of political trolls. The tweets data include the hashtags, so we calculate the number of hashtags in the 500 tweets for each account.

4) NUMBER OF URLS (URL)

Trolls use URLs to link tweets to other contents on Twitter or to external sites that have further information related to their agency's view. It is found that around 53% of the Russian trolls' tweets contain URLs, which is almost double the numbers of URLs in a sample of random tweets [31]. Trolls use URLs to increase their tweets incredibility [39]. Accordingly, the number of URLs can also be a good distinctive feature of political trolls. The tweets data in our dataset include URLs so we calculate the number of URLs in the 500 tweets for each account.

5) NUMBER OF RETWEETS (RET)

The political trolls do not work independently; they are members of a troll farm, so their works are always coordinated in order to amplify their effects. They retweet each other's tweets especially the tweets that were related to political events [13], [35], [40]. Boatwright *et al.* [13] analyzed strategies of the Russian trolls, based on the pattern of behaviors,

they found these trolls could be categorized to right trolls who posted right-leaning messages and left trolls who posted liberal messages. Around 76% of the tweets of left trolls and 60% of the right trolls are retweets. Trolls in each group mostly retweet from other trolls within the same group or from trolls that presented themselves as US local news aggregators. Using the flag that indicates if a tweet is original or retweet in our dataset, we count how many retweets among the last 500 tweets for each account.

6) NUMBER OF REPLY (REP)

In addition to retweets, reply is another approach used by political trolls to promote each other's accounts on Twitter as well as to disseminate propaganda. By analyzing the Russian trolls' tweets, it is found that around 20% of the Russian trolls' tweets were reply tweets [13]. Therefore, the percentage of replies among the tweets can be a good indicative feature of sponsored trolling. The tweets data in our dataset include a field called "in_reply_to_tweetid", so if that field contains data, it means this tweet is a reply. We used this to count the number of replies in the 500 tweets for each account.

7) PERCENTAGE OF WEEKENDS' TWEETS(WKD)

Keller *et al.* [23] analyzed trolls' behavior involved in the South Korean presidential election in 2012, they found temporal patterns, where trolls have significantly fewer posts during the weekends comparing to regular users who usually post more frequently during weekends. We found that is also valid for both Saudi and IRA trolls. Fig. 8 shows the average number of tweets per user each day of the week. It shows that both Saudi and IRA trolls have less average number of tweets on weekends comparing to ordinary users, taking into consideration that weekend in Saudi Arabia is Friday and Saturday, while weekend in Russia is Saturday and Sunday. Using the field "tweet_time" in tweets data, we calculate the percentage of tweets that were posted on weekends for each account.

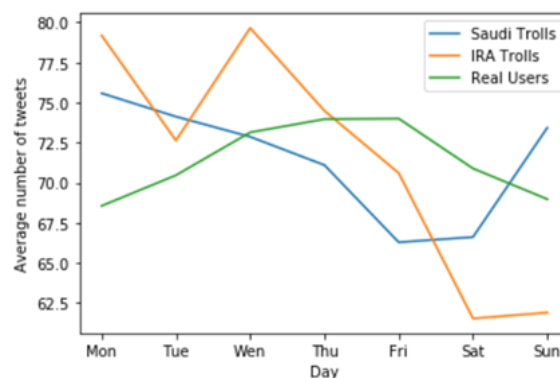


FIGURE 8. The average number of tweets per user in each day of the week.

8) TIME OF TWEETS (TIM)

Another aspect of temporal behavior patterns is the timing of tweets during the day. It is found that trolls commonly

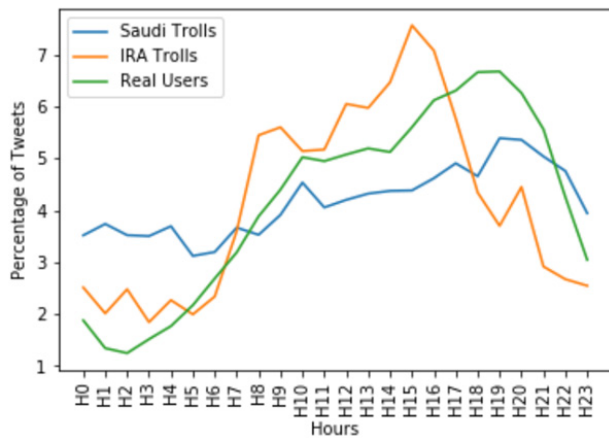


FIGURE 9. Percentage of tweets over the 24 hours.

tweet during regular office hours, while ordinary users post tweets more frequently after-work hours [23]. By analyzing the timing of tweets in our dataset, we also found variation in tweet timing between trolls and regular users as depicted by Fig. 9. It is found that trolls have more posts in the morning comparing to regular users, who post more frequently in the evening. Around 66% of regular users' tweets are posted between noon and midnight, where the Saudi trolls and IRA posted only 55%, and 59% respectively of their tweets at that time. Therefore, we calculate the ratio of tweets between 12:00 am and 12:00 pm for the 500 tweets for every account as a feature of behavior.

C. MACHINE LEARNING MODEL

We used scikit-learn [41], a framework of Python machine learning modules, to build four classifiers using the following approaches: decision tree, random forest, Adaboost, and gradient boost. First, we trained and tested these models on the Saudi trolls set to assess their performance using 10-fold cross-validation, where 80% of the data used for training and 20% for testing. Then we evaluated the performance of these models in identifying the IRA trolls without any training on this set.

1) DECISION TREE

The decision tree is one of the most popular supervised machine-learning approaches for classification, its output easily interpreted. It is based on tree structure that is constructed during data training by splitting training data recursively into subsets based on the values of the features. The leaf nodes of the tree represent the classes of the model. Each path from the root to the leaf nodes contains multiple internal nodes that represent the features used to make the decision, the branches are rules inferred during the training phase [42].

2) RANDOM FOREST

Random forest is a supervised machine learning approach developed by Breiman [43]. It is an ensemble model where multiple based models are combined in order to improve

the accuracy of the output, which is the aggregation of the individual models' results. Random forest is an extension of the bagging ensemble technique, it is composed of several decision trees, and training data is bootstrapped to create different datasets for each base model. It differs from bagged decision trees by using random subsets of features for each individual tree in order to reduce the correlation between features [44].

3) ADABOOST

Generally, boosting-based approaches aim to improve the performance of weak learners, which performed slightly better than random chance. In contrast to bagging approaches, which combine independent models and aggregate the outputs, boosting approaches combine weak models sequentially where each subsequent model refocuses on the observations that the previous one misclassified them. Adaboost, short for adaptive boosting, combines a series of weak learners, where initially all data instances are assigned equal weights, then after training a weak learner, the misclassified instances are assigned higher weights in order to make them more visible to be selected for training the next learner. This weight depends on the error value, i.e. the percentage of misclassified instances, the higher the error, the more is the weight assigned to these instances. This process is repeated until all instances are classified correctly or the specified number of estimators is reached. Moreover, each of the weak learners is assigned a weight based on its accuracy, so during the classification process, the models with high classifiers will have more impact on the final decision [45].

4) GRADIENT BOOSTING MACHINE (GBM)

Gradient Boosted Machine (GBM) is another model of boosting approaches, which combines multiple weak models to improve the overall performance. It is composed of multiple consecutive base models, typically decision trees. Each successive tree improves the prediction of the previous one by further minimize the loss function using gradient decent optimization algorithm [46]. The main difference between Adaboost and gradient boosting is how each one addresses the weakness of the previous learner, while Adaboost assigns a high weight for misclassified instances in training data, GBM use gradient decent algorithm to optimize the performance of the model.

IV. RESULTS

In this section, we discuss the results of using the above machine learning models for identifying sponsored-trolls. We start by evaluating the models on the Saudi troll accounts, then evaluate these models on IRA trolls.

A. TESTING THE MODELS ON THE SAUDI TROLLS

We evaluated the performance of the four machine learning classifiers on the Saudi trolls set in terms of accuracy, precision, recall, F-score, and Area Under the Curve (AUC). These metrics rely on the confusion matrix described in Table 2. The

TABLE 2. Confusion matrix.

	PREDICTED	TROLL	GENUINE
ACTUAL			
TROLL		TP	FN
GENUINE		FP	TN

True Positive (TP) is the number of troll accounts that are correctly identified by the model as trolls. The False Negative (FN) is the number of troll accounts that are misclassified by the model as genuine accounts. The False Positive (FP) is the number of genuine accounts that are misclassified by the model as trolls. The True Negative (TN) is the number of genuine accounts that are correctly classified by the model as genuine.

The accuracy is the percentage of the number of accounts that are classified correctly versus the total of all accounts, as shown in (1).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

The precision represents the percentage of the accounts classified correctly as trolls over the total accounts that are classified as trolls, as shown in (2).

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

The recall represents the percentage of the accounts classified correctly as trolls overall real troll accounts, as shown in (3).

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

F1-score represents the harmonic average of the precision and recall and calculated, as shown in (4).

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

Area Under the Curve (AUC) is a metric used with binary classification problems to assess the model performance, it is the summary of the Receiver Operator Characteristic (ROC) curve which is a probability curve plots true positive rates (recall) vs. false positive rate at all classification thresholds. The value of AUC ranges from 0 to 1, which indicates the ability of the classifier to distinguish between classes. A higher value of AUC means better performance of distinguishing between trolls and genuine accounts.

Table 3 shows the performance of the four models on the Saudi trolls. We can observe that all the four models have an excellent performance in identifying the Saudi sponsored-trolls in terms of all the metrics. Gradient boost model slightly outperforms the other three models, its accuracy reaches 94.4% and F-score reaches 0.942. Generally, the high performance of the four models in terms of all metrics shows that sponsored-trolls' behaviors on Twitter are distinguishable from ordinary users' behaviors.

Fig. 10 shows the performance of the four models across all possible classifications thresholds. The three classifiers,

TABLE 3. Performance of the models in identifying Saudi trolls.

Model	Accuracy	Precision	Recall	F-Score	AUC
Decision Tree	0.927	0.935	0.912	0.924	0.926
Random Forest	0.937	0.950	0.918	0.934	0.937
Adaboost	0.923	0.923	0.906	0.919	0.922
Gradient Boost	0.944	0.956	0.927	0.942	0.944

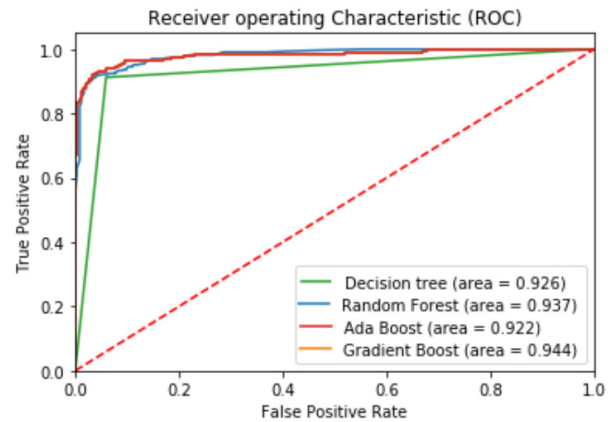


FIGURE 10. Area under the ROC curve for the four classifiers.

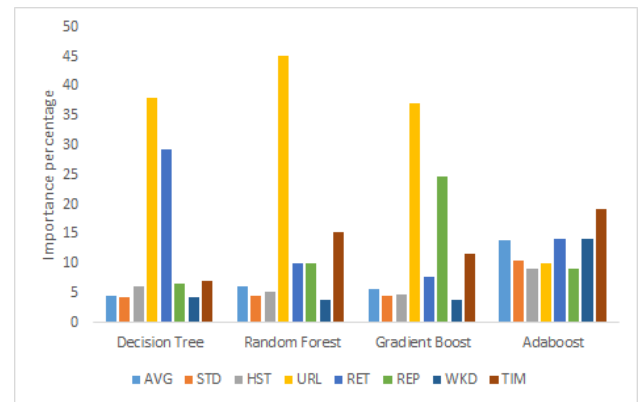


FIGURE 11. The features importance for the four classifiers.

random forest, Adaboost, and gradient boost, have dominated decision tree all the time.

Fig 11 illustrates the importance of the features in each one of the four models. The number of URLs is the most important feature in decision tree, random forest, and gradient boost with importance percentage 38%, 45%, and 37% respectively, while the time of the tweets is the most important feature when using Adaboost model with percentage 19%. The second important feature is different in each model; it is the number of retweet in decision tree (29.4%), tweets time in random forest (15.2%), number of replies in gradient boost (24.8%), and percentage of weekend tweets in Adaboost (14.2%). The third important feature is tweet time in both decision tree (7%) and gradient boost (11.5%), number of retweet in random forest (10%), and the average of tweets

TABLE 4. Performance of the models in identifying IRA trolls.

Model	Accuracy
Decision Tree	0.721
Random Forest	0.677
Adaboost	0.726
Gradient Boost	0.641

in Adaboost (13.9%). Accordingly, the features contribute differently based in each model.

B. TESTING THE MODELS ON THE IRA TROLLS

To assess our assumption, we evaluated the performance of the four models in identifying IRA accounts. We only use recall metric because the rest of the metrics depend on real accounts, however this set does not have real accounts. Thus, in this case, the accuracy is the same as recall.

Table 3 depicts the performance of the four models in identifying the trolls in IRA set. Decision tree model and Adaboost outperform random forest and gradient boost. Generally, this result shows that political trolls have common behaviors on social networks regardless of their languages or the organization that hired them. These behaviors are different from ordinary users, therefore machine-learning approaches that were trained on a set of these accounts are able to identify, to some extent, these troll accounts in general.

The difference in models' accuracy in identifying IRA trolls comparing to Saudi trolls might be due to different strategies used by each one of the two organizations, for example, we found that IRA trolls used more hashtags than Saudi trolls. On the other side, Saudi trolls used more replies than IRA trolls. Using more instances of real accounts from different countries and different cultures might improve the accuracy of the developed models.

V. DISCUSSION

The results showed that state-sponsored troll accounts on Twitter can be identified using machine learning algorithms based only on behavioral features. The results of testing the proposed classifiers on both the Saudi and IRA trolls support our initial assumption that sponsored-trolls in general are employees driven by extrinsic rewards, so they have common behaviors that cannot be hidden. In contrast to the previous works [34], [36], [37], our work does not use any linguistic features related to the contents that posted by the trolls, so it does not require any natural language processing. This helps to develop general classifiers to identify trolls in different farms. At the same time, our work has high performance similar to the performance in [34] and [36], and better than the performance in [37].

Our work contributes to fill the research gap in addressing this phenomenon by developing automated tools to identify political troll accounts in order to immune public against opinion manipulation. With automatic identification of troll

accounts, regular users can avoid these accounts or consume the information they post with caution. Social media platforms have huge data of their users' behaviors, so they can build more robust approaches to identify and suspend troll accounts quickly.

We limit the behavior features to the eight different user's activities on Twitter described in section three; however, political trolls are centrally coordinated, so more behavioral features that capture the relationships between trolls within the same farm will improve the accuracy of classification. Moreover, the developed classifiers were trained on the Saudi trolls only. Using more trolls from different farms for training will improve the performance.

VI. CONCLUSION AND FUTURE WORK

In this paper, we shed the light on an increasing phenomenon on social media, which is state-sponsored trolls. They are human operators hired by governments, or political organizations to support their political stance, or engage in foreign influence operations. Based on our initial assumption that state-sponsored trolls cannot completely hide their suspicious behaviors, we developed eight features of users' behaviors on Twitter, and used them with four machine-learning classifiers to identify state-sponsored trolls. The results showed that we can build general classifiers that can identify political trolls solely based on their behaviors regardless of the contents they post, or the organization they work for.

This work can be extended by using more behavioral features like how trolls engage with other users through following, and mentioning. Moreover, the dataset can be enriched more by using more genuine accounts from different countries, and different cultures, then assess the developed models against trolls from other countries and organizations that have been disclosed by Twitter lately.

REFERENCES

- [1] C. K. Jha and O. Kodila-Tedika, "Does social media promote democracy? Some empirical evidence," *J. Policy Model.*, vol. 42, no. 2, pp. 271–290, Mar. 2020.
- [2] P. Iosifidis and M. Wheeler, *Public Spheres and Mediated Social Networks in the Western Context and Beyond*. London, U.K.: Palgrave, 2016.
- [3] S. Bradshaw and P. N. Howard, "The global disinformation order 2019 Global inventory of organised social media manipulation," Univ. Oxford, Oxford, U.K., Tech. Rep., 2019, pp. 1–21. [Online]. Available: <https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/09/CyberTroop-Report19.pdf>
- [4] T. Lokot and N. Diakopoulos, "News bots: Automating news and information dissemination on Twitter," *Digit. Journalism*, vol. 4, no. 6, pp. 682–699, Aug. 2016, doi: [10.1080/21670811.2015.1081822](https://doi.org/10.1080/21670811.2015.1081822).
- [5] S. Bradshaw, "Troops, trolls and troublemakers: A global inventory of organized social media manipulation," Univ. Oxford, Oxford, U.K., Comput. Propaganda Res. Project, Working Paper 2017.12, 2017. [Online]. Available: <https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/89/2017/07/Troops-Trolls-and-Troublemakers.pdf>
- [6] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Detecting automation of Twitter accounts: Are you a human, bot, or cyborg?" *IEEE Trans. Dependable Secure Comput.*, vol. 9, no. 6, pp. 811–824, Nov. 2012, doi: [10.1109/TDSC.2012.75](https://doi.org/10.1109/TDSC.2012.75).
- [7] R. Gorwa and D. Guilbeault, "Unpacking the social media bot: A typology to guide research and policy," *Policy Internet*, vol. 12, no. 2, pp. 225–248, Jun. 2020, doi: [10.1002/poi3.184](https://doi.org/10.1002/poi3.184).

- [8] C. Hardaker, "Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions," *J. Politeness Res. Lang., Behav., Culture*, vol. 6, no. 2, pp. 215–242, Jan. 2010, doi: [10.1515/jplr.2010.011](https://doi.org/10.1515/jplr.2010.011).
- [9] J. Paavola, T. Helo, H. Jalonen, M. Sartonen, and A.-M. Huhtinen, "Understanding the trolling phenomenon," *J. Inf. Warf.*, vol. 15, no. 4, pp. 100–111, Sep. 2016. [Online]. Available: <https://www.jstor.org/stable/26487554>
- [10] A. Spruds. (2015). *Internet Trolling as a Hybrid Warfare Tool: The Case of Latvia*. NATO Strategic Communications Centre of Excellence, Riga, Latvia. [Online]. Available: <https://www.stratcomcoe.org/download/file/3345>
- [11] C.-P. Chiang, H.-Y. Chen, T.-M. Tsai, S.-H. Chang, Y.-C. Chen, and S.-J. Wang, "Profiling operations of cyber army in manipulating public opinions," in *Proc. 6th Int. Conf. Frontiers Educ. Technol.*, Jun. 2020, pp. 222–225, doi: [10.1145/3404709.3404766](https://doi.org/10.1145/3404709.3404766).
- [12] L. G. Stewart, A. Arif, and K. Starbird, "Examining trolls and polarization with a retweet network," in *Proc. ACM WSDM, Workshop Misinformation Misbehavior Mining Web*, 2018, pp. 1–6.
- [13] D. Linvill and P. Warren, "Troll factories: The internet research agency and state-sponsored agenda building," Resource Centre Media Freedom Eur., Working Paper, 2018. [Online]. Available: <https://www.rcmediafreedom.eu/Publications/Academic-sources/Troll-Factories-The-Internet-Research-Agency-and-State-Sponsored-Agenda-Building>
- [14] J. Lee, "Review of the mueller report," *J. Bus. Ethics*, vol. 163, no. 1, pp. 167–172, Apr. 2020, doi: [10.1007/s10551-019-04357-8](https://doi.org/10.1007/s10551-019-04357-8).
- [15] A. Badawy, A. Addawood, K. Lerman, and E. Ferrara, "Characterizing the 2016 russian IRA influence campaign," *Social Netw. Anal. Mining*, vol. 9, no. 1, pp. 1–11, Dec. 2019, doi: [10.1007/s13278-019-0578-6](https://doi.org/10.1007/s13278-019-0578-6).
- [16] A. Badawy, K. Lerman, and E. Ferrara, "Who falls for online political manipulation?" 2018, *arXiv:1808.03281*. [Online]. Available: <http://arxiv.org/abs/1808.03281>
- [17] C. A. Bail, "Assessing the Russian Internet Research Agency's impact on the political attitudes and behaviors of American Twitter users in late 2017," *Proc. Nat. Acad. Sci. USA*, vol. 117, no. 1, pp. 243–250, 2020, doi: [10.1073/pnas.1906420116](https://doi.org/10.1073/pnas.1906420116).
- [18] C. Llewellyn, L. Cram, A. Favero, and R. L. Hill, "Russian troll hunting in a brexit Twitter archive," in *Proc. 18th ACM/IEEE Joint Conf. Digit. Libraries*, May 2018, pp. 361–362, doi: [10.1145/3197026.3203876](https://doi.org/10.1145/3197026.3203876).
- [19] E. Ferrara, "Disinformation and social bot operations in the run up to the 2017 French presidential election," Jul. 2017, *arXiv:1707.00086*. [Online]. Available: <https://arxiv.org/abs/1707.00086>
- [20] M. O. Jones, "The gulf information war propaganda, fake news, and fake trends: The weaponization of Twitter bots in the gulf crisis," *Int. J. Commun.*, vol. 13, p. 27, Mar. 2019.
- [21] D. Boyd, S. Golder, and G. Lotan, "Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter," in *Proc. 43rd Hawaii Int. Conf. Syst. Sci.*, Jan. 2010, pp. 1–10, doi: [10.1109/HICSS.2010.412](https://doi.org/10.1109/HICSS.2010.412).
- [22] F. Guerrero-Solé, "Interactive behavior in political discussions on Twitter: Politicians, media, and citizens' patterns of interaction in the 2015 and 2016 electoral campaigns in Spain," *Social Media+ Soc.*, vol. 4, no. 4, Oct. 2018, Art. no. 205630511880877, doi: [10.1177/2056305118808776](https://doi.org/10.1177/2056305118808776).
- [23] F. B. Keller, D. Schoch, S. Stier, and J. Yang, "Political astroturfing on Twitter: How to coordinate a disinformation campaign," *Political Commun.*, vol. 37, no. 2, pp. 256–280, Mar. 2020, doi: [10.1080/10584609.2019.1661888](https://doi.org/10.1080/10584609.2019.1661888).
- [24] D. G. Lilleker and K. Koc-Michalska, "What drives political participation? Motivations and mobilization in a digital age," *Political Commun.*, vol. 34, no. 1, pp. 21–43, Jan. 2017, doi: [10.1080/10584609.2016.1225235](https://doi.org/10.1080/10584609.2016.1225235).
- [25] S. J. Grossman and O. D. Hart, "An analysis of the principal-agent problem," in *Foundations of Insurance Economics*. New York, NY, USA: Springer, 1992, pp. 302–340.
- [26] J. Bishop, "Dealing with Internet trolling in political online communities," *Int. J. E-Politics*, vol. 5, no. 4, pp. 1–20, Oct. 2014, doi: [10.4018/ijep.2014100101](https://doi.org/10.4018/ijep.2014100101).
- [27] P. B. O'Sullivan and A. J. Flanagin, "Reconceptualizing 'flaming' and other problematic messages," *New Media Soc.*, vol. 5, no. 1, pp. 69–94, Mar. 2003, doi: [10.1177/1461444803005001908](https://doi.org/10.1177/1461444803005001908).
- [28] M. A. Al-garadi, K. D. Varathan, and S. D. Ravana, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network," *Comput. Hum. Behav.*, vol. 63, pp. 433–443, Oct. 2016, doi: [10.1016/j.chb.2016.05.051](https://doi.org/10.1016/j.chb.2016.05.051).
- [29] P. Fornacciari, M. Mordonini, A. Poggi, L. Sani, and M. Tomaiuolo, "A holistic system for troll detection on Twitter," *Comput. Hum. Behav.*, vol. 89, pp. 258–268, Dec. 2018, doi: [10.1016/j.chb.2018.08.008](https://doi.org/10.1016/j.chb.2018.08.008).
- [30] T. Mihaylov and P. Nakov, "Hunting for troll comments in news community forums," 2019, *arXiv:1911.08113*. [Online]. Available: <http://arxiv.org/abs/1911.08113>
- [31] S. Zannettou, T. Caulfield, E. De Cristofaro, M. Sirivianos, G. Stringhini, and J. Blackburn, "Disinformation warfare: Understanding state-sponsored trolls on Twitter and their influence on the Web," in *Proc. Companion Proc. World Wide Web Conf.*, May 2019, pp. 218–226, doi: [10.1145/3308560.3316495](https://doi.org/10.1145/3308560.3316495).
- [32] A. Badawy, E. Ferrara, and K. Lerman, "Analyzing the digital traces of political manipulation: The 2016 russian interference Twitter campaign," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2018, pp. 258–265, doi: [10.1109/ASONAM.2018.8508646](https://doi.org/10.1109/ASONAM.2018.8508646).
- [33] E. Van Der Walt and J. Eloff, "Using machine learning to detect fake identities: Bots vs humans," *IEEE Access*, vol. 6, pp. 6540–6549, 2018, doi: [10.1109/ACCESS.2018.2796018](https://doi.org/10.1109/ACCESS.2018.2796018).
- [34] B. Ghanem, D. Buscaldi, and P. Rosso, "TexTrolls: Identifying russian trolls on Twitter from a textual perspective," 2019, *arXiv:1910.01340*. [Online]. Available: <http://arxiv.org/abs/1910.01340>
- [35] S. Zannettou, T. Caulfield, W. Setzer, M. Sirivianos, G. Stringhini, and J. Blackburn, "Who let the trolls out?" in *Proc. 10th ACM Conf. Web Sci. (WebSci)*, 2019, pp. 353–362, doi: [10.1145/3292522.3326016](https://doi.org/10.1145/3292522.3326016).
- [36] J. Im, E. Chandrasekharan, J. Sargent, P. Lighthammer, T. Denby, A. Bhargava, L. Hemphill, D. Jurgens, and E. Gilbert, "Still out there: Modeling and identifying russian troll accounts on Twitter," 2019, *arXiv:1901.11162*. [Online]. Available: <http://arxiv.org/abs/1901.11162>
- [37] L. Luceri, S. Giordano, and E. Ferrara, "Detecting troll behavior via inverse reinforcement learning: A case study of russian trolls in the 2016 US election," 2020, *arXiv:2001.10570*. [Online]. Available: <http://arxiv.org/abs/2001.10570>
- [38] A. Vesselkov, B. Finley, and J. Vankka, "Russian trolls speaking russian: Regional Twitter operations and MH17," in *Proc. 12th ACM Conf. Web Sci.*, Jul. 2020, pp. 86–95, doi: [10.1145/3394231.3397898](https://doi.org/10.1145/3394231.3397898).
- [39] A. Gupta and P. Kumaraguru, "Credibility ranking of tweets during high impact events," in *Proc. 1st Workshop Privacy Secur. Online Social Media (PSOSM)*, 2012, pp. 2–8, doi: [10.1145/2185354.2185356](https://doi.org/10.1145/2185354.2185356).
- [40] C. Llewellyn, L. Cram, R. L. Hill, and A. Favero, "For whom the bell trolls: Shifting troll behaviour in the Twitter brexit debate," *JCMS: J. Common Market Stud.*, vol. 57, no. 5, pp. 1148–1164, Sep. 2019, doi: [10.1111/jcms.12882](https://doi.org/10.1111/jcms.12882).
- [41] F. Pedregosa, G. Varoquaux, and A. Gramfort, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.
- [42] S. Kotsiantis, "Supervised machine learning: A review of classification techniques," *Informatica*, vol. 31, no. 3, pp. 249–268, 2007.
- [43] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
- [44] A. Criminisi, "Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning," *Found. Trends Comput. Graph. Vis.*, vol. 7, nos. 2–3, pp. 81–227, 2011, doi: [10.1561/06000000035](https://doi.org/10.1561/06000000035).
- [45] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Proc. 13th Int. Conf. Int. Conf. Mach. Learn.*, 1996, pp. 148–156.
- [46] J. H. Friedman, "Stochastic gradient boosting," *Comput. Statist. Data Anal.*, vol. 38, no. 4, pp. 367–378, Feb. 2002, doi: [10.1016/S0167-9473\(01\)00065-2](https://doi.org/10.1016/S0167-9473(01)00065-2).



SALEH ALHAZBI (Senior Member, IEEE) received the B.S. degree in computer science from the University of Mosul, Iraq, in 1996, the M.S. degree in computer science from New Mexico State University, USA, in 2001, and the Ph.D. degree in computer science from the University of Science, Malaysia, in 2009. His research interests include applied machine learning, software engineering, and computer science education.

• • •