Early Detection and Prevention of Malicious User Behavior on Twitter Using Deep Learning Techniques

Rubén Sánchez-Corcuera, Arkaitz Zubiaga, and Aitor Almeida

***Abstract*—Organized misinformation campaigns on Twitter continue to proliferate, even as the platform acknowledges such activities through its transparency center. These deceptive initiatives significantly impact vital societal issues, including climate change, thus spurring research aimed at pinpointing and intercepting these malicious actors. Present-day algorithms for detecting bots harness an array of data drawn from user profiles, tweets, and network configurations, delivering commendable outcomes. Yet, these strategies mainly concentrate on postincident identification of malevolent users, hinging on static training datasets that categorize individuals based on historical activities. Diverging from this approach, we advocate for a forward-thinking methodology, which utilizes user data to foresee and mitigate potential threats before their realization, thereby cultivating more secure, equitable, and unbiased online communities. To this end, our proposed technique forecasts malevolent activities by tracing the projected trajectories of user embeddings before any malevolent action materializes. For validation, we employed a dynamic directed multigraph paradigm to chronicle the evolving engagements between Twitter users. When juxtaposed against the identical dataset, our technique eclipses contemporary methodologies by an impressive 40.66% in F score (F1 score) in the anticipatory identification of harmful users. Furthermore, we undertook a model evaluation exercise to gauge the efficiency of distinct system elements.**

***Index Terms*—Foreseeing, malicious users, social networks, Twitter.**

I. Introduction

SINCE the creation of computers, urban environments and societies have experienced momentous shifts, incorporating technological advancements that have seamlessly integrated into daily routines [1]. Numerous civic processes have emerged or evolved to foster intelligent cities in this context. Examples include crowdsourcing initiatives [2], [3], [4], surveys [5], [6], and the integration of the Internet of Things (IoT) with other techniques [7], [8], [9]. The landscape of communication mediums has not been spared from this metamorphosis; it has seen profound evolutions, culminating in Online Social Networks as the preeminent communication channels in modern-day societies [10]. Yet, this advantageous shift has simultaneously attracted undesirable users who contaminate the digital social sphere, disseminating deceptive content for their objectives [11]. As recognized by Twitter through its TTC [12], the platform has increasingly found itself in the crosshairs of such malevolence, witnessing a surge of content disruptors and malignant users targeting lawful participants.

The occurrence of politically motivated attacks, such as the induction attack during the Brexit vote in the United Kingdom [13] and the revelation of 200 Russian accounts posing as American citizens attempting to influence the 2016 American elections [14], exemplify the societal consequences of such actions. However, attacks are not limited to political motivations. For instance, misinformation about climate change on Twitter [15], [16] highlights the spread of false information and conspiracy theories, which can have detrimental effects. Common forms of climate change misinformation on Twitter encompass denying the existence of climate change, disputing human responsibility, and disseminating inaccurate information about its impacts. Such misinformation undermines public trust in scientific evidence and hampers efforts to mitigate the consequences of climate change.

Detecting users involved in these attacks has become increasingly important, spurring research in this field. However, most existing work on bot and malicious user detection has followed a forensic perspective [17], [18], [19], conducting post hoc experiments using data from concluded events that have already been addressed. Previous approaches focus on detecting attacks retrospectively. In contrast, our proposed work takes a preventive approach by leveraging user data preceding events to predict and counteract attacks before they occur. Although the forensic approach benefits from more extensive data, our preventive scenario offers a realistic setting that exploits evolving data dynamics and adapts to behavioral changes over time. We operationalize this preventive scenario by retrieving and expanding sets of malicious users published by Twitter through its TTC initiative, thereby simulating a dynamic environment that utilizes limited user histories preceding the events.

Manuscript received 17 November 2023; revised 17 April 2024 and 17 June 2024; accepted 21 June 2024. Date of publication 12 July 2024; date of current version 2 October 2024. This work was supported by the Ministry of Economy, Industry and Competitiveness of Spain under Grant PID2021-128969OB-I00 (INCEPTION project). *(Corresponding author: Rubén Sánchez-Corcuera.)*

Rubén Sánchez-Corcuera and Aitor Almeida are with the Faculty of Engineering & DeustoTech, University of Deusto, 48007 Bilbo, Bizkaia, Spain (e-mail: ruben.sanchez@deusto.es).

Arkaitz Zubiaga is with the Social Data Science Laboratory, Queen Mary University of London, E1 4NS London, U.K.

Digital Object Identifier 10.1109/TCSS.2024.3419171

In recent years, social bot accounts have become increasingly sophisticated, making distinguishing them from legitimate users difficult, thanks to advancements in deepfakes and other deceiving algorithms [18]. This poses a challenge for traditional feature-based models, as they need to maintain their efficacy against these new malicious user models. Consequently, researchers in the field have expanded the feature set of detection models to include factors such as network topology of user relationships and activity patterns.

This manuscript presents a pioneering framework for detecting malicious users in a preventive manner by leveraging information from network topology, user interactions, and semantic features extracted from URLs and hashtags in their tweets. In our research, we define malicious users as individuals or entities utilizing their online presence to undermine digital ecosystems through various detrimental activities, each with unique characteristics, such as how they are controlled and their impacts on social platforms and their user communities. These activities include, but are not limited to the following.

1) *Disinformation Campaigns*: Coordinated efforts to create, distribute, and amplify factually incorrect or misleading information with the objective to manipulate public perception, erode trust in institutions, or sow discord among communities. This is particularly prevalent in contexts where the shaping of public opinion can have far-reaching consequences.
2) *Idea Induction Attempts*: The strategic and often covert promotion of specific ideologies, beliefs, or narratives intended to subtly influence or radically alter the cognitive and emotional landscapes of target audiences. Unlike disinformation campaigns, idea induction is characterized by its nuanced approach to embedding certain viewpoints within the discourse, thereby guiding audiences toward adopting these perspectives as their own, frequently observed in political influence, brand marketing, and social engineering.
3) *Social Engineering*: The manipulation of individuals into divulging confidential information or performing actions detrimental to themselves or others. Social engineering exploits human psychology rather than technical hacking techniques, leveraging tactics such as pretexting, baiting, and quid pro quo.

In our methodology, we employ the jointly optimizing dynamics and interactions for embeddings (JODIEs) [20] framework, a dynamic graph network model that is instrumental in predicting and understanding user behavior on social media platforms. JODIE's core functionality revolves around its ability to dynamically model user–item interactions over time, thus allowing for the prediction of future interactions. This is achieved by maintaining and updating embeddings for users and items based on their interactions, effectively capturing the evolving nature of user behavior. Specifically for the problem of malicious user detection, JODIE helps in foreseeing potential harmful activities by analyzing the projected trajectories of user embeddings. By integrating JODIE, we can proactively identify users likely to engage in malicious behavior before such actions materialize, leveraging this predictive capability to mitigate potential threats early. This anticipatory approach underpins our strategy for enhancing online security, illustrating JODIE's critical role in our solution to combat organized misinformation campaigns and malicious activities on Twitter.

The uniqueness of our approach lies in its proactive stance toward predicting malicious user behavior. Unlike traditional methods that primarily focus on postincident analysis based on static datasets, our technique emphasizes the early identification of potential threats by dynamically analyzing Twitter interactions. This forward-thinking methodology allows for the prevention of malicious activities before they occur, contributing to the enhancement of online safety and trustworthiness. To the best of our knowledge, this is the first model trained and tested in a preventive manner for this task.

To assess the performance and validity of our proposed model, we formulated two key research questions that are addressed in this study as follows.

1) *RQ 1*: Can malicious users be detected proactively by utilizing their social network interaction information?
2) *RQ 2*: How quickly can malicious users be identified once their attacks have commenced?

To investigate these research questions, we conducted several classification experiments to analyze the ability to preemptively classify malicious users and determine the number of interactions required to identify them. To create a dataset tailored explicitly for this task, we merged data from the TTC with data obtained from legitimate Twitter users. This approach ensures that the identification of malicious users is based on the labeling provided by the social network itself.

The article presents several key contributions as follows.

1) A novel methodology is proposed for training and testing preventive malicious user detection models, taking into account the temporal aspect of the data. Specifically, this approach leverages a dynamic graph-based model that updates user embeddings in real time, a departure from traditional static models. This enables the prediction of malicious activities by analyzing changes in user behavior patterns over time, offering a more proactive and accurate detection mechanism. Additionally, our technique integrates a new set of semantic features extracted from URLs and hashtags within tweets, enhancing the model's ability to understand and predict the context of potential malicious actions.
2) A unique approach utilizing Twitter interactions is introduced, employing advanced analytical techniques to dissect the nature of user interactions. By analyzing the sequence and context of tweets, retweets (RT), mentions, and replies, our model uncovers underlying patterns indicative of malicious intent. This method stands out by its ability to infer potential malicious activities from seemingly benign interactions, using a novel algorithm that assesses the risk based on interaction dynamics and content analysis.
3) A comprehensive dataset comprising 596 221 tweets from legitimate users is meticulously compiled, focusing on interactions related to malicious users identified in three

state-backed operations as reported by Twitter. This dataset is unparalleled in its integration of dynamic interaction data and semantic features, such as URLs and hashtags, providing a rich resource for understanding the tactics of malicious actors. The dataset's structure and content offer unprecedented insights into the behavior of malicious users, facilitating the development of more sophisticated detection models [21].

These contributions collectively enhance the understanding and detection of malicious users on social media platforms, offering insights into proactive measures to mitigate the dissemination of fake information and malicious activities.

II. Related Work

The identification and classification of malicious users on Twitter have gained significant attention due to the potential impact of their actions on societal processes and individuals' lives. Researchers have been actively developing models and automated systems to classify these users among legitimate Twitter users accurately [22]. This research endeavor has spanned over a decade, with numerous systems implemented using diverse methodologies for detecting malicious users [18].

The continuous efforts in this field reflect the ongoing importance of addressing the presence of malicious users and the evolving nature of their behaviors on social media platforms. By developing and refining detection methods, researchers aim to enhance online communities' overall integrity, trustworthiness, and security while mitigating the negative impacts caused by the dissemination of fake information, malicious activities, and attempts to manipulate public opinion.

In this section, we review the landscape of research focused on the detection and prevention of malicious user behavior on social media platforms, particularly Twitter. Our review is purposefully structured to underline three pivotal axes: historical methodologies in bot and malicious user detection, the evolution of deep learning techniques in enhancing detection precision, and the emerging importance of predictive rather than reactive approaches in combating online malice. Furthermore, as we employed a graph-based approach we also analyze models that specifically use this data structures to tackle malicious user detection problem.

This evolution from basic profile-based methods to sophisticated, anticipatory models underscores the field's shift toward proactive detection strategies. By highlighting these developments, our review not only showcases the gradual sophistication of detection mechanisms but also sets the stage for our proposed methodology, which aims to predict and prevent malicious activities before they unfold.

By organizing the related work in this manner, we aim to provide a comprehensive understanding of where our research fits into the ongoing conversation about online safety and integrity. This structure not only elucidates the progression of detection methods over time but also rationalizes the necessity of our forward-thinking approach in addressing the limitations of existing models.

A. Malicious Account Detection

In recent years, the detection of bots and malicious accounts on Twitter has emerged as a highly active field of research [18]. This growing interest has even led to competition to identify the best-performing detection models [23]. Researchers have also extended their studies to analyze the various attack vectors that malicious actors can exploit [24], [25]. While many published studies focus on extracting features from users or user-generated content to identify patterns associated with different types of users, the specific set of features and algorithms employed vary across studies.

Early works employed simple features and rule-based approaches for detecting malicious users [26]. However, these methods quickly became outdated due to the rapid sophistication of malicious users. Subsequently, more advanced models emerged, leveraging over a thousand features and employing various algorithms for classification [27]. As datasets and benchmarks evolved [28], [29], classification models became more sophisticated and adapted to new paradigms such as deep learning. For instance, some models based on long short-term memory (LSTM) networks have been proposed to analyze the content of users' tweets for classification [30]. This shift toward deep learning enables the analysis of unstructured information, including the network structures connected users create. Models utilizing graph convolutional networks (GCNs) have been introduced to exploit these user relationships [31].

Another critical aspect of malicious user detection models is how they employ data for training. Two approaches can be identified based on the work proposed by [32]. The first approach, known as the forensic approach, does not consider the temporality of the data. Therefore, the dataset can be fully utilized without any limitations. In contrast, the preventive approach respects the temporal order of the data by defining a point in the dataset that represents the present time. Data occurring after this point remains unseen by the model until the testing phase.

The forensic approach allows models to gather more information about users after an attack, improving their ability to detect subsequent attacks. On the other hand, the preventive approach reflects the reality of social networks more accurately, as it imposes temporal constraints on the analysis performed before classifying users. This approach enables the development of models capable of detecting malicious users in preattack stages and preventing legitimate users from falling victim to them.

B. Dynamic Graphs

The utilization of dynamic graphs in our study is motivated by the intrinsic nature of social media interactions, which are inherently temporal and evolve over time. Unlike static graphs, which provide a snapshot of user interactions at a single point in time, dynamic graphs allow us to capture the sequential and temporal aspects of user behavior and relationships. This capability is crucial for understanding and predicting the actions of malicious users on platforms such as Twitter, where strategies and behaviors may quickly evolve to evade detection. Dynamic graphs enable the modeling of changes in user interactions, network structures, and community formations, facilitating the

TABLE I
COMPARISON OF BOT DETECTION DATASETS CREATED BY THE SCIENTIFIC COMMUNITY AND OURS WITH THE PROPOSED REQUIREMENTS FOR OUR TASK

Dataset Name	Malicious Users	Legit Users	Users Related	Timeline Format	Annotated by Twitter	Malicious Users
Ours	**1594**	**2736**	✓	✓	✓	✓
Caverlee-2011 [38]	15 483	14 833	✓	✓		
Cresci-2015 [17]	1950	1950		✓		
Cresci-2017 [39]	7049	2764		✓		
Varol-2017 [40]	733	1495	✓			
Gilani-2017 [41]	1090	1413		✓		
Cresci-stock-2018 [42]	7102	6174	✓	✓		
Midterm-2018 [43]	42 446	8092	✓			
Pronbots-2019 [43]	17 882	-	✓	✓		
Celebrity-2019 [43]	-	5918	✓	✓	✓	
Vendor-purchased-2019 [43]	1087	-		✓		
Botometer-feedback-2019 [43]	143	380		✓		
Political-bots-2019 [43]	62	-	✓	✓		
Cresci-rt-bust-2019 [44]	353	340	✓	✓		
Botwiki-2019 [28]	698	-		✓		
Verified-2019 [28]	-	1987		✓	✓	
Kaiser [45]	27	1048	✓		✓∼	
Astroturf [27]	505	-				
Twibot-20 [29]	6589	5237	✓∼			

early identification of anomalous patterns indicative of malicious activity. By employing dynamic graphs, our approach can adapt to and preempt emerging threats, leveraging temporal information to predict future interactions and behaviors of users with high precision. This temporal aspect is key to our predictive model, as it allows for a more nuanced and accurate identification of potential malicious users before they engage in harmful activities, enhancing the effectiveness of our detection mechanism.

Dynamic graphs have gained significant attention in recent years, and several notable works on node classification can be adapted for user classification and, consequently, malicious user detection. DyRep [33] introduced a model that encodes temporal information into node representations by considering the evolving structure of the graph. It incorporates a temporal point process-based attention mechanism focusing on the destination node's neighborhood. TGAT [34] employs a self-attention mechanism and a time encoding technique. By stacking multiple TGAT layers, the model can recognize node embeddings as functions of time and infer embeddings for unseen nodes as the graph evolves. TGN [35] currently represents the state of the art for node classification and link prediction in dynamic graphs. It combines memory modules to store long-term information and a graph-based embedding module to generate up-to-date node embeddings. Jodie [20] uses a time projection embedding module to predict embedding trajectories over time. The model utilizes two recurrent neural networks (RNNs) to update the representations of different entities (e.g., users and actions) in different positions at different times. This feature allows the model to identify users employing various strategies, such as mention spamming [36]. We have chosen Jodie as the model for our experiments because it enables the update of user and action representations, which can be advantageous in identifying users following different strategies during an attack. In Section V, we will compare the performance of Jodie and TGN in the task of preventive bot classification.

III. DATA GATHERING AND SOURCES

The detection of bots, content polluters, and malicious users has accumulated a significant amount of data over the years, which researchers have utilized to train and validate their detection models. The observatory on social media (OSoMe) at Indiana University is a prominent institution that maintains a comprehensive repository of bot detection datasets [37]. While OSoMe offers a diverse range of datasets, it is important to note that the institution does not collect them. The repository includes datasets collected by various researchers and organizations, each with its format, objectives, and information.

For this study, we analyzed the most well-known datasets in the scientific community and examined their respective characteristics.

Upon evaluating the datasets developed by the scientific community for bot detection, we determined that most of them did not satisfy our specific requirements for achieving our objectives. Table I compares the users included in each dataset and their alignment with the aforementioned objectives. We have outlined our requirements as follows.

1) The dataset must encompass both malicious and legitimate users.
2) Malicious and legitimate users within the dataset should be associated with or collected during the same period.
3) The dataset must include the captured tweets of the users, respecting the temporal nature of the publications and adhering to Twitter's broadcasting rules.
4) User labels, particularly for malicious users, must be validated by Twitter through account suspension or the application of automated user labels.
5) The dataset should encompass various types of malicious users, including bots, cyborgs, and human-operated accounts.

Given these reasons, we opted to create a new dataset incorporating official data from suspended Twitter users involved in state-backed coordinated operations, as identified and

acknowledged by the social network. Additionally, we collected data from legitimate users associated with the identified malicious users, enabling the creation of a comprehensive dataset featuring both legitimate and malicious users operating within the same circumstances. This dataset was constructed using the TTC data, which will be further elaborated upon in the subsequent section.

A. TTC

The Twitter Transparency Report was established in 2012 to inform users about government pressures, including censorship requests and demands for user information or content removal. Over time, Twitter transformed the transparency report into a dedicated website called the TTC. This platform encompasses various sections that outline Twitter's commitment to transparency, with a particular focus on the Security & Integrity section, which highlights the social network's efforts to enhance protection and authenticity.

Within the security & integrity section, Twitter addresses Information Operations, which refer to state-linked campaigns involving inauthentic influence tactics carried out by multiple users, often with the knowledge and support of their respective governments. Since 2018, Twitter has been releasing data on detected operations through the TTC, allowing academia, journalists, and other interested parties to analyze and investigate these activities, as well as develop preventive systems against such attacks. Every few months, Twitter publishes a blog post on the TTC, summarizing the operations detected since the last report. These blog posts provide details about the identified operations, their targets, and the number of suspended users. Alongside each post, Twitter also releases a dataset containing information on the accounts involved in the operations and their associated tweets.

In the datasets related to state-backed operations, Twitter anonymized the information of less well-known users (those with fewer than 5000 followers) to minimize potential harm or damage that may occur in case of false positives during an attack. However, researchers specializing in Twitter can request a fully unhashed version of the data for research purposes only.

As of June 2022, the transparency portal of Twitter has published information on 42 coordinated actions. Given the significant number of actions, their diverse nature, and the resource-intensive process of collecting data from legitimate users, we have selected three distinct ones originating from China, Iran, and Russia. Each one possesses unique motives and varies in terms of scale and impact as follows.

1) *China (July 2019):* During the 2019 Hong Kong Protests, a massive movement against the proposed extradition law amendment, Twitter identified 936 accounts originating from within the People's Republic of China (PRC) that aimed to instigate political unrest in Hong Kong. These accounts were involved in discrediting the legitimacy and political positions of the protest movement. Since Twitter is blocked in China, these accounts utilized VPNs or specific unblocked IP addresses to access the platform. Notably, these 936 accounts were part of a larger network comprising over 200 000 accounts dedicated to similar activities.
2) *Iran (June 2019):* Twitter uncovered 1666 accounts linked to the Iranian government. These accounts collectively published nearly 2 million tweets with a bias that favored Iran's diplomatic and geostrategic perspectives on global news. As engaging in platform manipulation violates Twitter's rules, the accounts were subsequently suspended.
3) *Russia (May 2020):* Twitter detected 1152 accounts associated with a media website called Current Policy, which engages in state-backed political propaganda within Russia. These accounts were involved in coordinated activities such as cross posting and amplifying content in an inauthentic manner. Their objectives included promoting the United Russia Party and attacking political dissidents.

After conducting a manual inspection of the datasets, we made several observations. While previous research has often referred to accounts involved in attacks and network contamination as bots, our TTC datasets analysis revealed that not all banned users in information operations can be classified as bots or fake users. As a result, we propose categorizing these users as malicious users and developing classification systems that can identify them regardless of how their accounts are managed. Therefore, throughout this work, we group such users under the category of malicious users.

Furthermore, we observed that the number of accounts mentioned in the TTC blog posts needed to correspond to the number of users provided in the datasets in some cases. However, in other cases, while the total number of detected or suspended accounts matched, not all had tweets available in the declared data. Hence, this study reports the number of users and tweets available in the downloaded unhashed datasets.

To collect legitimate users, we decided to select users who had connected with the malicious users but had not been flagged by Twitter as malicious users. Many of the attacks carried out on Twitter use hashtag-based strategies such as *hashtag flooding* or *hashtag hijacking* [46] to carry out their main attack. Therefore, we decided to use the most commonly used hashtags by malicious users to identify legitimate users. We grouped all the hashtags used in the tweets from the provided dataset and selected the ones that represented 80% of the total hashtags. This approach allowed legitimate users to connect with malicious users through these hashtags. To accomplish this, we used a crawler tool to collect information from users who tweeted with the selected hashtags. Subsequently, we employed another crawler to retrieve all the tweets posted by legitimate users within a 48-hour window before and after the coordinated action. To manage the collection process efficiently, especially for prolific users with a large number of tweets, we decided to limit the number of tweets per legitimate user to 1000. The tweets from legitimate users are available [21].

Finally, to merge the data of malicious and legitimate users, we created a dynamic graph where the interactions between users served as edges, and the users themselves represented the nodes, as depicted in Fig. 1. The primary forms of user

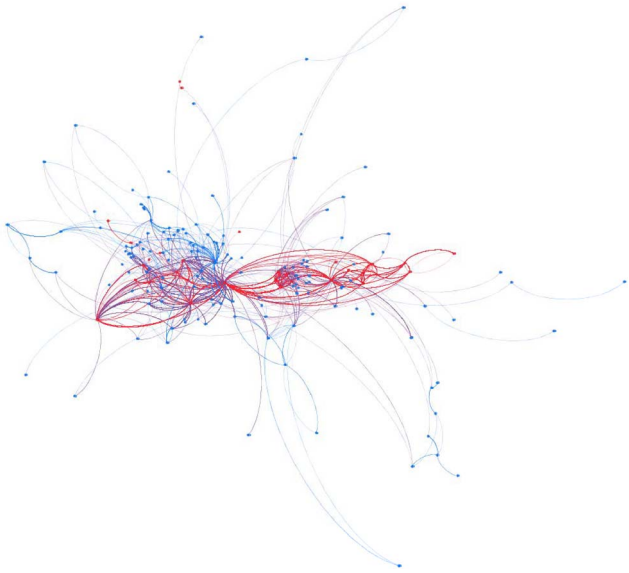

Fig. 1. Detailed network model illustrating the interactions among users within a social media platform. This model highlights the pathways through which information (or misinformation) flows between nodes (users), where nodes are color-coded by user type (blue for legit users and red for malicious ones) and edges represent different types of interactions (e.g., tweets, RTs, and mentions). The directional arrows indicate the flow of communication, emphasizing the influence patterns that are critical for identifying malicious behavior.

TABLE II
NODE AND EDGE STATISTICS OF THE DATASETS

	China	Iran	Russia
#Legit nodes	192 434 (344)	96 762 (684)	123 488 (1527)
#Malicious nodes	191	389	974
#Edges	2 029 418	1 868 433	1 142 663

Note: The number in brackets refers to the number of nodes as being creators of interactions and, therefore, classified.

interaction on Twitter, such as RT, mentions, and replies, were utilized to establish the edges. Since misleading hashtags or URLs are prevalent in Twitter attacks, we encoded them as features of the edges to leverage them with the foreseeing model. The statistics for each dataset are provided in Tables II and III.

IV. PROPOSED MODEL

We employed JODIE and a classification algorithm to detect malicious users on Twitter in advance. The JODIE model [20] is particularly suitable for the detection of malicious users due to its ability to dynamically capture and analyze the temporal sequence of user interactions. This feature is essential for our research as it allows for the identification of evolving patterns of malicious behavior that are not apparent in static snapshots of user data. By continuously updating user embeddings, JODIE provides a real-time perspective on user behavior, enabling the early detection of actions that deviate from typical user patterns, which could indicate malicious intent. It has previously been applied to predict user interactions and subreddits in Reddit's online social network. We incorporated a final classifier that determines whether users are likely to become malicious based on the predicted interactions to enhance the detection process.

TABLE III
BREAKDOWN OF TWEETS IN THE DATASETS BY TYPE OF INTERACTION AND USER

User Type	Interaction Type	China	Iran	Russia
Legit	Mention	26.32%	19.97%	3.70%
	Reply	19.45%	25.46%	7.35%
	RT	2.00%	2.28%	0.79%
	Plain tweets	52.23%	52.28%	88.16%
	Total tweets	321 812	557 750	1 489 737
Malicious	Mention	6.46%	3.12%	1.34%
	Reply	16.40%	2.51%	13.96%
	RT	39.12%	38.33%	12.08%
	Plain tweets	38.02%	56.04%	72.62%
	Total tweets	1 707 606	1 310 683	3 441 965

In the context of social networks, where user interactions are not only frequent but also varied across different vectors (e.g., tweets, retweets, and mentions), JODIE's capacity to process and learn from these interactions in real-time positions it as an ideal framework for anticipating and mitigating potential threats before they manifest into larger security concerns.

We developed a two-stage model to facilitate embedding foreseeing and embedding classification, as illustrated in Fig. 3. Each stage is responsible for a specific task: the first stage focuses on predicting the subsequent user–item interactions using JODIE, while the second stage classifies users as either malicious or legitimate based on their representations derived from the previous step. The details of these models are explained in the following sections.

A. Embedding Foreseeing Model (EFM)

The JODIEs model, proposed by Kumar et al. [20], is designed to capture the sequential interactions between users and items in various domains, such as e-commerce and social networks. The model utilizes representation learning by creating embeddings for users and items in Euclidean space and analyzing their trajectories based on their interactions over time. The implementation of JODIE involves using two RNNs, one responsible for modifying the embeddings of users and the other for item embeddings. One of the key features of the JODIE model is its ability to forecast future trajectories of users and items. To achieve this, the authors introduce a projection operation that learns to estimate the embeddings at any future time. This allows the model to capture users' and items' evolving behavior and properties over time.

In JODIE, users (U) and items (I) are represented using two types of embeddings: static embeddings and dynamic embeddings. Static embeddings remain unchanged over time and represent the long-term interests of users. The authors employ one-hot vectors to represent static embeddings, as they have been found to yield similar results to more complex embedding techniques. On the other hand, dynamic embeddings capture the characteristics of users and items at specific points in time and

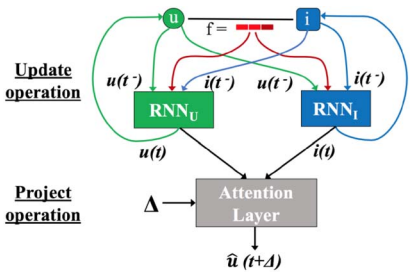

Fig. 2. Update and project operations of the JODIE model. Figure extracted from [20] for explanatory purposes.

are capable of evolving over time. The sequence of dynamic embeddings is referred to as the trajectory and forms the basis for modeling the temporal dynamics of user–item interactions. By leveraging the trajectory of dynamic embeddings, JODIE enables the prediction of future interactions between users and items, making it a suitable model for forecasting users' behaviors in sequential interaction scenarios.

JODIE proposes two operations that update and project the embeddings after each interaction, as shown in Fig. 2.

1) *Update Operation*: In the update operation, an interaction $S = (u, i, t, f)$ between a user u and an item i at a given time t, with its features f is used to generate the dynamic embeddings of that user and item at that specific time. The model uses two RNNs, called RNN_U and RNN_I; each one is in charge of updating the embeddings of users or items, respectively. Authors designed these RNNs as mutually recursive, i.e., when an interaction is processed at time t, the RNN_U updates the dynamic embedding of the user by using the previous embedding of that user $u(t^-)$ and item $i(t^-)$ right before the time t and regardless which has been its last interaction and the same for updating the items. The update operation for users and items is defined as

$$\begin{aligned}\mathbf{u(t)} &= \sigma(W_1^u u(t^-) + W_2^u i(t^-) + W_3^u f + W_4^u \Delta_u)\\ \mathbf{i(t)} &= \sigma(W_1^i i(t^-) + W_2^i u(t^-) + W_3^i f + W_4^i \Delta_i)\end{aligned} \quad (1)$$

where Δ denotes the time since the previous interaction of a user or item depending on its subindex (Δ_u or Δ_i). f is the feature vector of the interaction and $W_1^u ... W_4^u$ and $W_1^i ... W_4^i$ are trainable matrices of RNN_u and RNN_i, respectively. The authors justify using RNN instead of LSTMs, GRUs or other similar architectures because they experiment with them and obtained similar or worse results with those models that include more parameters.

2) *Projection Operation*: This operation is in charge of projecting the embedding of a user in the Euclidean space to the desired time. With this operation, the model can predict the trajectory of a user and, thus, predict the next item with which that user will interact. The dynamic embedding of a user u at a specific time t plus the elapsed time is necessary for this operation. First, they convert the time to a time-context vector by passing it through a linear layer. Then, the projected embedding is obtained by conducting an elementwise product of that time-context vector with the previous embedding of the user. More formally

$$\hat{u}(t + \Delta) = (1 + w) * u(t) \quad (2)$$

where $1 + w$ represents a temporal attention vector to scale the embedding of the user in the past. Thus, when $\Delta = 0$, the time vector $w = 0$ and the projected embedding is the same as in the time t. The larger the value of Δ, the more difference between the original embedding and the one projected as the time difference is more significant.

The authors of JODIE train the model to predict the next item a user will interact with. To train both update and projection operations, the authors decided to train the model using the projected embedding of the user $u(t + \Delta)$. One important decision made by Kumar et al. was to directly output a projected item embedding, $\tilde{j}(t + \Delta)$, instead of a probability of the interaction between the user and the items. Thus, when conducting the forward pass of the prediction layer, the JODIE model outputs the predicted item embedding, and the item with the closest embedding to the projected user embedding is returned as the next interaction.

Thus, JODIE model is trained to minimize the L_2 difference between the predicted item embedding $\tilde{j}(t + \Delta)$ and the real item embedding $[\bar{j}, j(t + \Delta^-)]$ as the concatenation of the original embedding of item j and it embedding immediately before the time Δ. This difference is calculated as follows: $||\tilde{j}(t + \Delta) - [\bar{j}, j(t + \Delta^-)]||_2$. For the prediction of future item embedding, the authors propose to employ the projected user embedding $\hat{u}(t + \Delta)$ the embedding of the previous item that the user has interacted with $i(t + \Delta^-)$ at the prior time to the interaction Δ (that is what the superscript—means). Finally, the static embeddings of both user ($\bar{u}$) and item ($\bar{i}$) are used in the fully connected layer used for the prediction as follows:

$$\tilde{j}(t + \Delta) = W_1 \hat{u}(t + \Delta) + W_2 \bar{u} + + W_3 i(t + \Delta^-) + W_4 \bar{i} + B \quad (3)$$

where $W_1^u ... W_4^u$ and the bias B make the linear layer.

As mentioned before, JODIE is trained to minimize the L_2 distance between the predicted item and the ground truth item in the interactions. However, the loss applied to the training phase includes two terms to prevent the dynamic embeddings of users and items from varying too much and also. These terms are scaled using λ_U and λ_I to ensure that losses are in the same range. The loss is calculated as follows:

$$\begin{aligned}\text{Loss} = &\sum_{(u,j,t,f) \in S} ||\tilde{j}(t + \Delta) - [\bar{j}, j(t + \Delta^-)]||_2 \\ &+ \lambda_U ||u(t) - u(t^-)||_2 \\ &+ \lambda_I ||j(t) - j(t^-)||_2.\end{aligned} \quad (4)$$

The original model is trained using what authors call *t-batches*, special batches developed to maintain temporal dependencies between interactions. To maintain these temporal dependencies, the authors create these t-batches following these

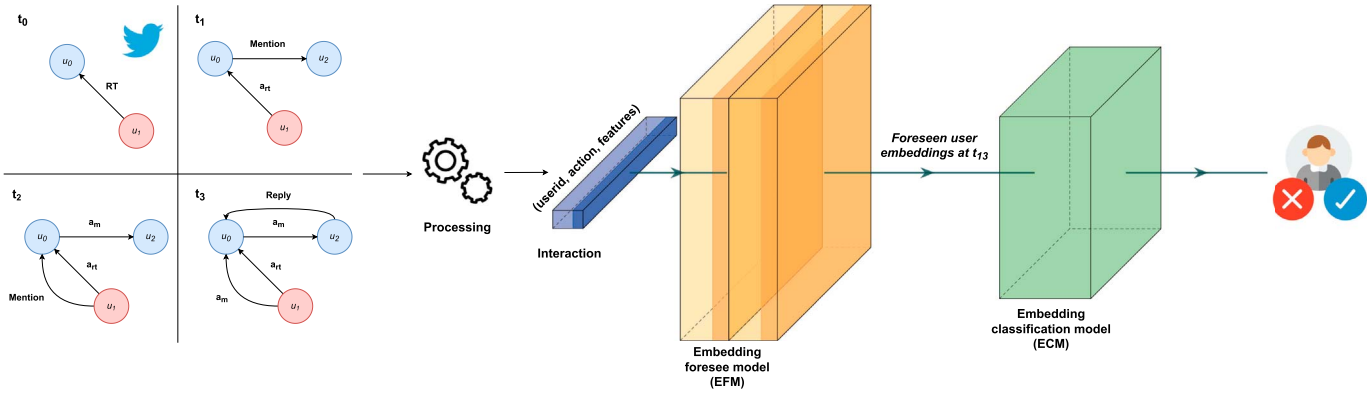

Fig. 3. Methodology of our approach, from creating the dynamic multigraph to classifying the embeddings at the desired point of the dataset. In the leftmost part of the figure, we can see a representation of how the dynamic multigraph is created. In *time 1* t_1, the *user 1* u_1 makes an RT to the *user 2* u_2, and the first edge between them is created. The actions follow one after the other, increasing the subindex of t and thus creating the temporality of the graph. This dynamic graph will later be processed into a list of interactions sorted by date.

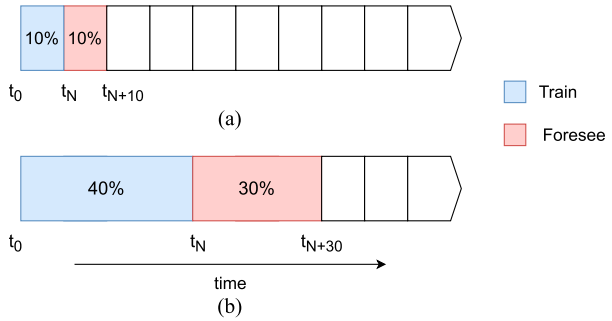

Fig. 4. Examples of dataset divisions used in our research. (a) Training of the model occurs using 10% of the dataset (t_N), along with a foreseeing size comprising 10% of the data (t_{N+10}). (b) Model undergoes training with 40% of the dataset (t_N) and with a foreseeing size of 30% of the data (t_{N+30}).

two requirements: 1) all interactions in a batch should be processed in parallel; and 2) the batches should be processed in increasing order to maintain the temporal ordering of the interactions. Thus, these two requirements may be summarized in that two interactions of the same batch do not share any common user or item. The authors prove that this batching system makes JODIE 9.2 times faster than its closest state-of-the-art model called *DeepCoevolve* [47].

In their study, Kumar et al. compared the performance of the JODIE model against six state-of-the-art models from three categories: deep recurrent recommender models, dynamic co-evolution models, and temporal network embedding models. The aim was to demonstrate the superiority of JODIE in various tasks and characteristics, including runtime and robustness. To evaluate its performance, the authors conducted five experiments to showcase the model's effectiveness compared to the existing approaches.

In contrast to the previous work by [20], which focused on using source and target users to predict future interactions, our model leverages Twitter interactions. This allows us to predict users' future actions, enabling us to identify and classify them as potentially malicious users before they complete their attacks. By analyzing the patterns and strategies exhibited by users in their Twitter interactions, we can proactively detect and classify malicious behavior, providing an opportunity for early intervention and mitigation.

B. Embedding Classification Model (ECM)

The ECM is responsible for classifying the embeddings generated by the embedding foreseeing model. In our implementation, we utilize the random forest (RF) classifier, as proposed by Breiman [48], with 300 estimators from the Scikit-learn library [49]. We conducted a model selection study where various algorithms were evaluated, and based on the results obtained across the three datasets, random forest emerged as the top-performing classifier. The results of the model selection study can be found in Table VII.

C. Methodology

In this section, we will explain the methodology we employed to train and test our model using the data extracted from the TTC. Figs. 3 and 4 illustrate the complete methodology used in our experiments and provide insights into how the data are divided for training and testing purposes. We will now outline and describe the steps involved in the methodology:

The methodology employed for training and testing our model using the TTC dataset can be summarized as follows.

1) *Conversion to Time-Ordered Interactions:* We process the dynamic multigraph derived from the TTC dataset and convert it into a time-ordered series of interactions $(t_1, t_2, ..., t_X)$ between users. Each interaction includes the origin and destination users, as well as a list of features. In our experiments, the features consist of tokenized versions of URLs and hashtags extracted from the associated tweets. In Fig. 3, the interactions are represented as edges between users, labeled as a_{RT} or a_m depending on the action conducted.
2) *Training Dataset Definition:* We define the training portion of the dataset by selecting a specific moment in time that represents the present. All the interactions that occurred before this moment are considered past data,

which will be unseen by the model and used for predicting future interactions. For example, in Fig. 4(a), the training dataset comprises the first 10% of the interactions. We denote the final timestamp of the training portion as t_N. Interactions up to t_N are utilized to train the *embedding foreseeing model*, specifically the embedding projection layer, to create user representations after foreseeing their future interactions.

3) *Selection of Last Timestamp for Foreseeing:* We determine the last timestamp for which we want to foresee user interactions, i.e., how far into the future we aim to predict interactions. In Fig. 4(a), the testing dataset consists of interactions occurring after t_N until 20% of the total interactions in the dataset have taken place. This timestamp in the dataset is denoted as t_{N+10}. The *embedding foreseeing model* is employed to predict the embeddings of users who have interactions between these defined timestamps.
4) *ECM:* Finally, the *embedding foreseeing model* produces user embeddings at t_{N+10}, which are then passed to the *ECM*. This model utilizes 100 random splits to train and classify the embeddings generated by the preceding model.

The application of this model to Twitter data allows us to train it with historical data of the desired users up to the current time. Consequently, we can predict users' future actions and classify their intentions before they carry out their attacks.

V. Foreseeing Malicious Users by Their Actions

In this section, we introduce a language-agnostic model aimed at proactively detecting malicious users through their interactions. We will provide an overview of the experimental setup, describe the dataset utilized, and present the results obtained. This setup directly tackles the research questions posed in our study.

A. Experimental Setup

In our evaluation, we compare our proposed approach with TGN, the state-of-the-art model for node classification tasks in dynamic graphs [35]. Temporal graph networks (TGNs) are an advanced class of models designed to handle dynamic graph data by maintaining and updating node representations over time. This method integrates a memory module that stores the historical interactions of each node, allowing the model to capture temporal dependencies and changes in graph structure effectively. TGN processes dynamic graphs by leveraging memory modules, which store state representations of nodes and capture changes over time. It uses a message-passing architecture where each node accumulates messages from its interactions, which are then used to update its state in the memory. This mechanism allows TGN to maintain updated node embeddings that reflect recent interactions and structural changes. The choice of TGN as a benchmark for comparison in our study stems from its innovative approach to handling temporal information and its proven effectiveness in diverse applications involving dynamic networks, such as social network analysis, recommendation systems, and anomaly detection. Its ability to update node embeddings continuously and capture temporal patterns provides a robust framework for predictive tasks in dynamic environments. By comparing our model with TGN, we aim to highlight the enhancements and specific contributions of our approach to the field of malicious user detection on social platforms.

As stated, TGN utilizes batches for faster training than other models. However, a drawback of batch processing is that node representations are not updated until the batch is complete, potentially resulting in the use of outdated node information. To address this issue, the authors propose a message and memory aggregation system that considers all interactions within a batch before generating node representations. Nevertheless, we argue that this approach may overlook timing information related to user actions when aggregating batch messages. Additionally, we describe the model selection study that supports our choice of models and demonstrates their performance.

During the development of our model, we conducted a model selection study to determine the best algorithms for both the EFM and the ECM. We performed multiple evaluations with a 10% foresee size, gradually moving t_N up to 50% of the dataset. To avoid using large percentages of data in the training sets, we limited the evaluation to 50% of the dataset interactions when selecting models that achieved better user classification. Additionally, one of our research questions (RQ 2) focused on the speed of detecting malicious users while they are conducting attacks. Therefore, we aimed to use the algorithm that required the least amount of data for classification, as it would provide faster results.

For the ECM study, we employed the following classification algorithms: support vector classifier (SVC) [50], multilayer perceptron (MLP) [51], K-nearest neighbors (KNNs) [52], and RF [48]. All algorithms were implemented using the Scikit-learn library with default hyperparameters. The results of this study are presented in Table VII.

After determining the best algorithm for the ECM, we evaluated the performance of the EFM. To achieve this, we modified the original MLP proposed in TGN to classify the embeddings using the selected algorithm from the previous analysis. The results obtained from TGN are compared with our approach in Table VIII.

Subsequently, we compared our approach to TGN in the foreseeing classification task to validate its effectiveness. By comparing our approach to the state-of-the-art model in a similar task, we aimed to answer RQ 1 and determine if it is possible to detect malicious users based on their interactions proactively.

To conduct the evaluation, we performed a series of user classification experiments using the three datasets described in Section III-A and different foresee sizes: 10%, 30%, and 50% of the total interactions in the dataset. In each iteration, we kept the foreseeing percentage fixed for the evaluation set and increased t_N by 10% until the dataset's final timestamp. For instance, Fig. 4(a) represents the first experiment in the 10% foreseeing evaluation series, while Fig. 4(b) represents the fourth experiment in the 30% foreseeing series.

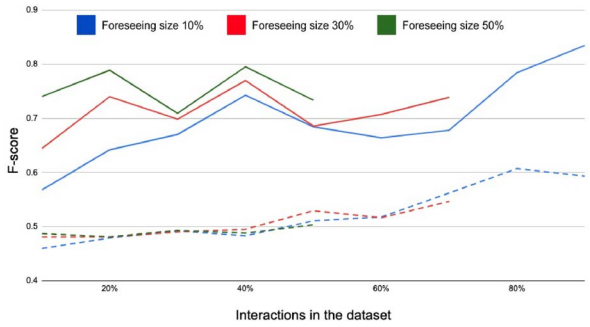

Fig. 5. Average of the results obtained in the three datasets for both models and with different foresee sizes. The solid lines represent the evaluation of our approach, and the dashed lines represent the evaluation of TGN.

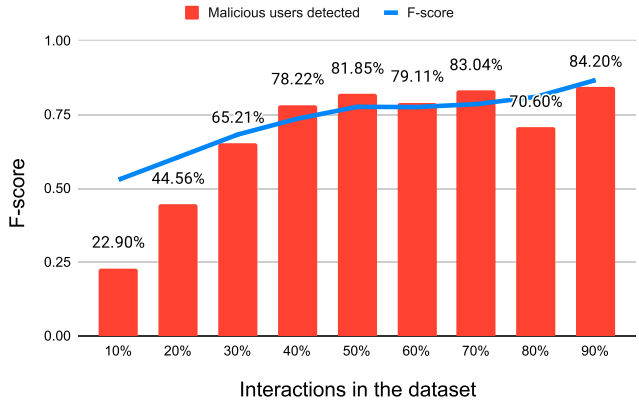

Fig. 6. Evaluation series conducted with a foreseeing size of 10% in the Iran dataset. The red bars represent the percentage of malicious users detected by the model, and the blue line represents the F-score obtained.

Our approach and the TGN model utilize their default ECMs. In TGN, the authors propose a custom MLP [51] classifier. However, we modified the training and testing process of this MLP to suit our experimental setup. As described in Section IV-C, the classifier is trained after the embedding projection to prevent data leakage. The train and test splits used for embedding classification are identical for both models.

All the evaluations conducted in our study utilized the following hyperparameters: an embedding size of 128, a learning rate of $1e^{-3}$, and a weight decay of $1e^{-5}$. These hyperparameters were consistent across all proposed evaluations. These hyperparameters were optimized by running a grid search and cross validation through the wandb tool [53]. Regarding the TGN model, we used the default hyperparameters as proposed in [35]. All evaluations were performed on an NVIDIA RTX 8000 graphics card.

B. Discussion

In a direct comparison against established methodologies, our approach demonstrates a substantial enhancement in the anticipatory identification of harmful Twitter users. Specifically, our technique achieves a remarkable 40.66% improvement in F score (F1 score) over contemporary strategies such as TGN. Fig. 5 illustrates the significant difference in average results, with our model consistently achieving more than 20 points higher than TGN. The evaluation across the three datasets shows that our model performs well, except for the Russian dataset, where it struggles after the midpoint of the dataset in all evaluation series. This performance decline could be attributed to a sudden change in the behavior of malicious and legitimate users, which the model was not trained to anticipate, failing to project accurate user embeddings.

This significant leap in performance underscores the efficacy of our forward-thinking model, which capitalizes on dynamic analysis of user interactions on Twitter to predict potential malicious activities. By leveraging a Dynamic Directed Multigraph paradigm for encapsulating evolving user interactions, our model not only excels in early detection but also sets a new benchmark in the domain of social media security. The ability to foresee and mitigate threats before their update not only enhances the resilience of digital communities against misinformation campaigns but also paves the way for more secure, equitable, and unbiased online environments.

The tables presented demonstrate that providing the model with minimal data does not yield good user representations, resulting in classification similar to that of a random classifier. However, as more interactions are included in the training set, our model achieves classification results with an F score above 0.75 before reaching 50% of t_N in some datasets. These results validate that the model does not require extensive training to begin projecting meaningful representations that the proposed algorithms can successfully classify. Furthermore, Fig. 6 shows an increasing accuracy in detecting malicious users over time, as represented by the red bars. With only 40% of the dataset interactions, the model can detect over 75% of the malicious users.

We employed the F score (F1 score) because it provides a more balanced view of model performance, especially useful in scenarios where class distributions are imbalanced. The F1 score is particularly effective in highlighting the effectiveness of our model in distinguishing between classes when both false positives and false negatives carry significant costs.

We also analyzed the foreseen size and its impact on the subsequent classification of embeddings. The evaluations were performed with foreseeing sizes of 10%, 30%, and 50% of the total interactions in the dataset. The results presented in Tables IV–VI demonstrate that our model achieves similar performance with the three different foresee sizes. However, it is evident that as the foresee size increases, the model achieves better results faster. Nonetheless, larger foresee sizes also make the model more susceptible to errors affecting its subsequent classification.

During the development of our model, we conducted a model selection study to determine the best algorithms for both the EFM and the ECM. The evaluations were performed with a 10% foresee size and t_N ranging from 10% to 50% of the dataset. We limited the evaluations to 50% of the dataset interactions to select models that achieved better user classification without requiring a large portion of the data for training.

TABLE IV
F-SCORE RESULTS FOR OUR APPROACH AND TGN ON THE THREE PROPOSED DATASETS USING A FORESEEING SIZE OF 10%

t_N	*China*		*Iran*		*Russia*	
	Our Approach	**TGN**	**Our Approach**	**TGN**	**Our Approach**	**TGN**
10%	**0.6142**	0.4720	**0.5930**	0.4220	**0.6742**	0.4880
20%	**0.6109**	0.4198	**0.6187**	0.4783	**0.6493**	0.5480
30%	**0.6527**	0.4308	**0.6456**	0.5086	**0.6809**	0.5441
40%	**0.6566**	0.4162	**0.701**	0.5398	**0.6244**	0.5025
50%	**0.766**	0.4677	**0.7426**	0.5131	**0.5915**	0.5558
60%	**0.7871**	0.4417	**0.7836**	0.5096	0.5832	**0.6172**
70%	**0.7698**	0.5143	**0.7697**	0.5519	0.5451	**0.6263**
80%	**0.8006**	0.5305	**0.8256**	0.6116	**0.7219**	0.6909
90%	**0.8414**	0.5593	**0.6892**	0.5740	**0.8248**	0.6511

Note: The bold values are mean the highest values for those lines.

TABLE V
F-SCORE RESULTS FOR OUR APPROACH AND TGN ON THE THREE PROPOSED DATASETS USING A FORESEEING SIZE OF 30%

t_N	*China*		*Iran*		*Russia*	
	Our Approach	**TGN**	**Our Approach**	**TGN**	**Our Approach**	**TGN**
10%	**0.6103**	0.424	**0.6485**	0.507	**0.7053**	0.518
20%	**0.6269**	0.406	**0.7033**	0.539	**0.6066**	0.509
30%	**0.7183**	0.436	**0.7428**	0.548	**0.6544**	0.494
40%	**0.6544**	0.415	**0.8041**	0.564	**0.6244**	0.519
50%	**0.7869**	0.486	**0.6905**	0.505	0.5999	**0.604**
60%	**0.8427**	0.455	**0.7874**	0.48	0.5719	**0.631**
70%	**0.8882**	0.496	**0.8329**	0.526	0.5137	**0.625**

Note: The bold values are mean the highest values for those lines.

TABLE VI
F-SCORE RESULTS FOR OUR APPROACH AND TGN ON THE THREE PROPOSED DATASETS USING A FORESEEING SIZE OF 50%

	China		*Iran*		*Russia*	
	Our Approach	**TGN**	**Our Approach**	**TGN**	**Our Approach**	**TGN**
10%	**0.7053**	0.432	**0.7446**	0.542	**0.7243**	0.494
20%	**0.7294**	0.403	**0.807**	0.562	**0.6114**	0.492
30%	**0.7157**	0.457	**0.8052**	0.542	**0.6684**	0.485
40%	**0.6547**	0.45	**0.8082**	0.475	**0.6227**	0.544
50%	**0.8965**	0.468	**0.5279**	0.457	0.5417	**0.597**

Note: The bold values are mean the highest values for those lines.

For the ECM study, we employed several classification algorithms, including SVC [50], MLP [51], KNN [52], and RF [48]. These algorithms were implemented using the Scikit-learn library with default hyperparameters. The results of the evaluations conducted in this study can be found in Table VII.

After selecting the best algorithm for the ECM, we focused on studying the EFM. To achieve this, we modified the MLP originally proposed in TGN to classify the embeddings, incorporating the selected algorithm from the previous evaluations. The results obtained by TGN were compared with our approach in Table VIII.

The results of the first study indicated that the different algorithms could classify the embeddings generated by the forecasting model with good results. However, considering the importance of early and accurate user classification for a preventive model, we selected the algorithm that achieved the best results in the initial sections of the dataset. In our case, the RF algorithm was chosen due to its favorable performance. While other algorithms may achieve better results at specific points in the dataset, the difference compared to RF, except for the Russian dataset, needed to be more significant to justify changing the classification algorithm.

The results presented in Table VIII indicate that our approach outperforms TGN and the best embedding classifier in two of the three datasets. This demonstrates the superiority of our EFM in accurately representing users by updating the representations of actions and users after each interaction. However, our model falls short compared to TNG+RF in the Russia dataset, showing a slight performance gap in the first four evaluations. Moreover, as observed in previous results, its performance deteriorates significantly when reaching the middle of the dataset. Nevertheless, based on the overall results, we can confidently assert that our proposed model is superior as it outperforms TGN in two of the three datasets considered.

VI. ETHICAL CONSIDERATIONS

The development of models to ensure social networks' security and neutrality raises critical ethical considerations. One of the key aspects is freedom of expression. Amnesty International and similar institutions define freedom of expression as the right to express, disseminate, seek, receive, and share information and ideas without fear of censorship. In line with this definition, social networks should uphold neutrality and allow users to express their ideas without limitations or fear of censorship. However, this concept of unlimited tolerance encounters the tolerance paradox, as Karl Popper articulated. The paradox states that if we extend unlimited tolerance even to those who are intolerant, it may lead to the destruction of a tolerant society and tolerance itself. Therefore, social networks need to strike a balance by limiting hate messages and actions from intolerant individuals to prevent the paradox of tolerance from occurring.

Another ethical dilemma arises when determining which users should be considered malicious and subject to censorship based on their behavior on the social network. In most datasets created by the scientific community, the researchers established the criteria for defining bots or malicious users. However, when using data provided by social networks such as Twitter through the TTC, the platform sets the criteria. Both cases highlight the challenge of relying on a "jury" to define the criteria for identifying malicious users. To address this, Twitter proposed the creation of the Twitter Moderation Research Consortium (TMRC), which comprises academia, civil society, NGOs, and journalism entities to study platform governance issues. Our work has been recognized through this initiative, and we have been invited to contribute.

The above considerations emphasize the need to design and implement models that combat the pollution of social networks and the spread of hatred toward others' ideas. Collaborative moderation models and processes such as the TMRC are essential for avoiding biases and ensuring a fair and healthy social network environment. These efforts aim to establish social networks as reliable information platforms instead of sources of problems that impact people's daily lives and even lead to psychological issues.

However, it is crucial to acknowledge that the proposed approach's accuracy could be more flawless. In a real environment, flagging users erroneously is possible, potentially resulting in a negative experience for those who are wrongfully

TABLE VII
F SCORE OF THE MODEL SELECTION STUDY CONDUCTED TO THE ECM WITH THE THREE DATASETS

	China				*Iran*				*Russia*			
t_N	**KNN**	**MLP**	**RF**	**SVC**	**KNN**	**MLP**	**RF**	**SVC**	**KNN**	**MLP**	**RF**	**SVC**
10%	0.5676	0.5621	**0.5930**	0.5534	0.5818	0.6000	0.5294	**0.6311**	0.5832	0.5812	**0.5841**	0.5829
20%	0.6092	0.5876	**0.6298**	0.5859	0.6202	0.6343	0.6057	**0.6455**	0.6631	0.6759	**0.6928**	0.6849
30%	0.6295	0.6400	0.6421	**0.6433**	0.6649	0.6497	**0.6825**	0.6788	0.6662	0.6287	**0.6869**	0.6325
40%	0.6782	0.6446	**0.6963**	0.6438	0.6963	0.7062	**0.7360**	0.7141	**0.8013**	0.7451	0.7990	0.7213
50%	0.6896	0.6673	**0.7192**	0.6757	0.7344	0.7310	**0.7762**	0.7380	**0.6125**	0.5789	0.5737	0.6054

Note: The foreseeing size was set to 10%, and the t_N was increased by 10% after each iteration. The bold values are mean the highest values for those lines.

TABLE VIII
F SCORE OF THE MODEL SELECTION STUDY CONDUCTED TO THE EFM WITH THE THREE DATASETS

	China		*Iran*		*Russia*	
t_N	**Our Approach**	**TGN+RF**	**Our Approach**	**TGN+RF**	**Our Approach**	**TGN+RF**
10%	0.5930	**0.6193**	**0.5294**	0.5040	0.5841	**0.5879**
20%	**0.6298**	0.5359	**0.6057**	0.5759	0.6928	**0.7022**
30%	**0.6421**	0.5769	**0.6825**	0.6455	0.6869	**0.7191**
40%	**0.6963**	0.6081	**0.7360**	0.7083	0.7990	**0.8128**
50%	**0.7192**	0.6654	**0.7762**	0.7682	0.5737	**0.8575**

Note: We changed employed the best-performing ECM for both approaches. The foreseeing size was set to 10%, and the t_N was increased by 10% after each iteration. The bold values are mean the highest values for those lines.

banned. Therefore, when applying this model to live Twitter data, it is necessary to conduct a manual analysis to identify possible false positives. Additionally, future work should focus on identifying the strategies employed by malicious users and forming groups to minimize the impact of their attacks, particularly when large numbers of users amplify them. The data provided by the TTC can offer valuable insights into the existence of such user groups.

VII. CONCLUSION AND FUTURE WORK

In this work, we have introduced a novel approach for preemptively detecting malicious users on social networks. Our approach leverages user interactions and features extracted from URLs and hashtags in tweets to identify malicious users. The model can detect malicious users by identifying their malicious actions or by recognizing early actions that resemble patterns observed in previously identified malicious users. Notably, our methodology differs from existing models by incorporating temporal patterns in the training process, allowing the model to capture the evolving nature of user behavior.

To validate our approach, we have applied state-of-the-art techniques for node classification in dynamic graphs to the problem of preemptive malicious user detection. In Section V, we have demonstrated the effectiveness of our proposed model by achieving competitive performance on the provided benchmark datasets. Additionally, we have observed that the model can project user embeddings over more extended time frames without significantly compromising subsequent classification results. Furthermore, we have conducted experiments to evaluate various classification algorithms for our model and selected the one that performed best on the benchmark datasets.

For future work, we propose focusing on detecting coordinated movements or communities of malicious users with similar characteristics. Detecting such communities or coordinated movements on Twitter could be a more effective approach than solely focusing on individuals, which can be challenging. It would also be valuable to analyze malicious users' strategies to carry out their attacks. By creating a taxonomy of these strategies and analyzing their behaviors, the model could generalize across different countries and identify patterns specific to malicious users from various regions. Furthermore, a deep analysis of the reason behind the abrupt reduction of the model's accuracy in some points of the datasets (e.g., 80% of Fig. 6) will be beneficial to understand the model's limitations and improve its performance. Finally, it would be beneficial for the security of social networks to develop a model that detects the point at which users transition from normal users to attackers. To create such a model, it would be necessary to have preattack data on users that is not available in the TTC or obtainable through suspending Twitter accounts.

REFERENCES

[1] R. Sánchez-Corcuera et al., "Smart cities survey: Technologies, application domains and challenges for the cities of the future," *Int. J. Distrib. Sensor Netw.*, vol. 15, no. 6, 2019, Art. no. 1550147719853984.
[2] X. Kong, X. Liu, B. Jedari, M. Li, L. Wan, and F. Xia, "Mobile crowdsourcing in smart cities," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 8095–8113, Oct. 2019.
[3] A. S. El-Wakeel, J. Li, A. Noureldin, H. S. Hassanein, and N. Zorba, "Towards a practical crowdsensing system for road surface conditions monitoring," *IEEE Internet Things J.*, vol. 5, no. 6, pp. 4672–4685, Dec. 2018.
[4] F. Corno, T. Montanaro, C. Migliore, and P. Castrogiovanni, "SmartBike: An IoT crowd sensing platform for monitoring city air pollution," *Int. J. Elect. Comput. Eng.*, vol. 7, no. 6, pp. 3602–3612, 2017.
[5] G. White, A. Zink, L. Codecá, and S. Clarke, "A digital twin smart city for citizen feedback," *Cities*, vol. 110, 2021, Art. no. 103064.
[6] F. R. Cecconi, N. Moretti, M. C. Dejaco, S. Maltese, and L. C. Tagliabue, "Community involvement in urban maintenance prioritization," in *Proc. AEIT Int. Annu. Conf.*, Piscataway, NJ, USA: IEEE, 2017, pp. 1–6.
[7] T. Montanaro, I. Sergi, M. Basile, L. Mainetti, and L. Patrono, "An IoT-aware solution to support governments in air pollution monitoring based on the combination of real-time data and citizen feedback," *Sensors*, vol. 22, no. 3, 2022, Art. no. 1000.
[8] L. C. Tagliabue, F. R. Cecconi, S. Maltese, S. Rinaldi, A. L. C. Ciribini, and A. Flammini, "Leveraging digital twin for sustainability assessment of an educational building," *Sustainability*, vol. 13, no. 2, p. 480, 2021.
[9] S. Rinaldi, F. Bittenbinder, C. Liu, P. Bellagente, L. C. Tagliabue, and A. L. C. Ciribini, "Bi-directional interactions between users and cognitive buildings by means of smartphone app," in *Proc. IEEE Int. Smart Cities Conf. (ISC2)*, Piscataway, NJ, USA: IEEE, 2016, pp. 1–6.
[10] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *Proc. 19th Int. Conf. World Wide Web*, 2010, pp. 591–600.

[11] S. Volkova and J. Y. Jang, "Misleading or falsification: Inferring deceptive strategies and types in online news and social media," in *Companion Proc. Web Conf.*, 2018, pp. 575–583.

[12] I. Twitter, "Twitter transparency center," Sep. 2021. Accessed: Jun. 17, 2023. [Online]. Available: https://transparency.twitter.com/

[13] M. T. Bastos and D. Mercea, "The Brexit botnet and user-generated hyperpartisan news," *Social Sci. Comput. Rev.*, vol. 37, no. 1, pp. 38–54, 2019.

[14] A. Bovet and H. A. Makse, "Influence of fake news in Twitter during the 2016 US presidential election," *Nature Commun.*, vol. 10, no. 1, pp. 1–14, 2019.

[15] L. Bode and E. K. Vraga, "In related news, that was wrong: The correction of misinformation through related stories functionality in social media," *J. Commun.*, vol. 65, no. 4, pp. 619–638, 2015.

[16] S. Van der Linden, A. Leiserowitz, S. Rosenthal, and E. Maibach, "Inoculating the public against misinformation about climate change," *Global Challenges*, vol. 1, no. 2, 2017, Art. no. 1600008.

[17] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "Fame for sale: Efficient detection of fake Twitter followers," *Decis. Support Syst.*, vol. 80, pp. 56–71, 2015.

[18] S. Cresci, "A decade of social bot detection," *Commun. ACM*, vol. 63, no. 10, pp. 72–83, 2020.

[19] S. Farooqi and Z. Shafiq, "Measurement and early detection of third-party application abuse on Twitter," in *Proc. WWW*, New York, NY, USA: ACM, 2019. [Online]. Available: https://doi.org/10.1145/3308558.3313515

[20] S. Kumar, X. Zhang, and J. Leskovec, "Predicting dynamic embedding trajectory in temporal interaction networks," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2019, pp. 1269–1278.

[21] R. Sánchez-Corcuera, "rubensancor/foreseeing_actions_data: v1.0.0," 2022. [Online]. Available https://doi.org/10.5281/zenodo.6862544

[22] S. Yardi et al., "Detecting spam in a Twitter network," *First Monday*, vol. 15, no. 1, 2010.

[23] V. S. Subrahmanian et al., "The DARPA Twitter bot challenge," *Computer*, vol. 49, no. 6, pp. 38–46, 2016.

[24] R. Sánchez-Corcuera, A. Bilbao-Jayo, U. Zulaika, and A. Almeida, "Analysing centralities for organisational role inference in online social networks," *Eng. Appl. Artif. Intell.*, vol. 99, 2021, Art. no. 104129.

[25] R. Sánchez-Corcuera, A. Zubiaga, and A. Almeida, "Analyzing the existence of organization specific languages on Twitter," *IEEE Access*, vol. 9, pp. 111463–111471, 2021.

[26] B. Wang, A. Zubiaga, M. Liakata, and R. Procter, "Making the most of tweet-inherent features for social spam detection on Twitter," 2015, *arXiv:1503.07405*.

[27] M. Sayyadiharikandeh, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer, "Detection of novel social bots by ensembles of specialized classifiers," in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manage.*, 2020, pp. 2725–2732.

[28] K.-C. Yang, O. Varol, P.-M. Hui, and F. Menczer, "Scalable and generalizable social bot detection through data selection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 1096–1103.

[29] S. Feng, H. Wan, N. Wang, J. Li, and M. Luo, "Twibot-20: A comprehensive Twitter bot detection benchmark," 2021, *arXiv:2106.13088*.

[30] S. Kudugunta and E. Ferrara, "Deep neural networks for bot detection," *Inf. Sci.*, vol. 467, pp. 312–322, 2018.

[31] S. Ali Alhosseini, R. Bin Tareaf, P. Najafi, and C. Meinel, "Detect me if you can: Spam bot detection using inductive representation learning," in *Companion Proc. World Wide Web Conf.*, 2019, pp. 148–153.

[32] D. F. Milon-Flores and R. L. Cordeiro, "How to take advantage of behavioral features for the early detection of grooming in online conversations," *Knowl.-Based Syst.*, vol. 240, 2022, Art. no. 108017.

[33] R. Trivedi, M. Farajtabar, P. Biswal, and H. Zha, "DyRep: Learning representations over dynamic graphs," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–25.

[34] D. Xu, C. Ruan, E. Korpeoglu, S. Kumar, and K. Achan, "Inductive representation learning on temporal graphs," 2020, *arXiv:2002.07962*.

[35] E. Rossi, B. Chamberlain, F. Frasca, D. Eynard, F. Monti, and M. Bronstein, "Temporal graph networks for deep learning on dynamic graphs," 2020, *arXiv:2006.10637*.

[36] C. Besel, J. Echeverria, and S. Zhou, "Full cycle analysis of a large-scale botnet attack on Twitter," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Piscataway, NJ, USA: IEEE, 2018, pp. 170–177.

[37] OSoMe, "Bot repository," 2022. Accessed: Jun. 17, 2023. [Online]. Available: https://botometer.osome.iu.edu/bot-repository

[38] K. Lee, B. Eoff, and J. Caverlee, "Seven months with the devils: A long-term study of content polluters on Twitter," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 5, 2011, pp. 185–192.

[39] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race," in *Proc. 26th Int. Conf. World Wide Web Companion*, 2017, pp. 963–972.

[40] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini, "Online human-bot interactions: Detection, estimation, and characterization," in *Proc. 11th Int. AAAI Conf. Web Social Media*, 2017, pp. 280–289.

[41] Z. Gilani, R. Farahbakhsh, G. Tyson, L. Wang, and J. Crowcroft, "Of bots and humans (on Twitter)," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, 2017, pp. 349–354.

[42] S. Cresci, F. Lillo, D. Regoli, S. Tardelli, and M. Tesconi, "$FAKE: Evidence of spam and bot activity in stock microblogs on Twitter," in *Proc. 12th Int. AAAI Conf. Web Social Media*, 2018, pp. 580–583.

[43] K.-C. Yang, O. Varol, C. A. Davis, E. Ferrara, A. Flammini, and F. Menczer, "Arming the public with artificial intelligence to counter social bots," *Human Behav. Emerg. Technol.*, vol. 1, no. 1, pp. 48–61, 2019.

[44] M. Mazza, S. Cresci, M. Avvenuti, W. Quattrociocchi, and M. Tesconi, "RTbust: Exploiting temporal patterns for botnet detection on twitter," in *Proc. 10th ACM Conf. Web Sci.*, 2019, pp. 183–192.

[45] A. Rauchfleisch and J. Kaiser, "The false positive problem of automatic bot detection in social science research," *PLoS One*, vol. 15, no. 10, 2020, Art. no. e0241045.

[46] N. Jain, P. Agarwal, and J. Pruthi, "Hashjacker-detection and analysis of hashtag hijacking on Twitter," *Int. J. Comput. Appl.*, vol. 114, no. 19, pp. 17–20, 2015.

[47] H. Dai, Y. Wang, R. Trivedi, and L. Song, "Deep coevolutionary network: Embedding user and item features for recommendation," 2016, *arXiv:1609.03675*.

[48] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[49] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[50] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[51] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, USA: JMLR, 2010, pp. 249–256.

[52] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *Amer. Statistician*, vol. 46, no. 3, pp. 175–185, 1992.

[53] L. Biewald, "Experiment tracking with weights and biases," wandb.com, 2024. Accessed: Jun. 17, 2023. [Online]. Available: https://www.wandb.com/