Russo-Ukrainian War: Prediction and explanation of Twitter suspension

Alexander Shevtsov
University of Crete (CSD-UOC)
Technical University of Crete (TUC)
Heraklion, Greece
asevtsov@tuc.gr

Despoina Antonakaki
Technical University of Crete (TUC)
ICS-FORTH
Chania, Greece
dantonakaki@tuc.gr

Ioannis Lamprou
University of Crete (CSD-UOC)
Technical University of Crete (TUC)
Chania, Greece
ilamprou1@tuc.gr

Ioannis Kontogiorgakis
Technical University of Crete (TUC)
Chania, Greece

Polyvios Pratikakis
ICS-FORTH
Heraklion, Greece

Sotiris Ioannidis
Technical University of Crete (TUC)
Chania, Greece

Abstract—On 24 February 2022, Russia invaded Ukraine, starting what is now known as the Russo-Ukrainian War, initiating an online discourse on SNs. Twitter one of the most popular SNs, with an open and democratic character, enables a transparent discussion among its large user base. Unfortunately, this often leads to Twitter's policy violations, propaganda, abusive actions, civil integrity violations, and consequently to user accounts' suspension and deletion. This study focuses on the Twitter suspension mechanism and the analysis of shared content and features leading to an accurate machine-learning suspension prediction. Toward this goal, we have obtained a dataset containing 107.7M tweets, originating from 9.8 million users, using Twitter API. We extract the categories of shared content of the suspended accounts and explain their characteristics, through the extraction of text embeddings in junction with cosine similarity clustering. Our results reveal scam campaigns taking advantage of trending topics regarding the Russia-Ukrainian conflict for Bitcoin and Ethereum fraud, spam, and advertisement campaigns. Additionally, we apply a ML methodology including a SHapley Additive explainability model to understand and explain how user accounts get suspended.

Index Terms—Twitter, user suspension, ML, explainability, SHAP, Russo-Ukrainian war

I. INTRODUCTION

On 24th February 2022, Russia's invasion of Ukraine, also known as the Russo-Ukrainian War, triggered extensive discussions on social media. Twitter, being one of the most prominent SNs, provides an open and democratic platform for transparent discussions among its vast user base. However, this openness can lead to policy violations, propaganda, abusive actions, and violations of civil integrity, resulting in the suspension and deletion of user accounts. This study focuses on understanding the mechanism behind Twitter suspension and analyzing the shared content and user account features that may lead to such suspensions.

To achieve this goal, we obtained a dataset comprising 107.7M tweets originating from 9.8M users, collected via Twitter API. We analyze the content shared by suspended accounts and investigate their characteristics using text embeddings and cosine similarity clustering. Our findings reveal the presence of scams and advertisement campaigns exploiting trending topics related to the Russia-Ukraine conflict towards Bitcoin and Ethereum fraud. Additionally, we employ various ML classification models and a SHapley Additive explainability [1] model to comprehend and explain the factors contributing to user account suspensions.

In response to the Ukraine war, Twitter has announced new moderation policies to reduce the amplification of Russian government accounts and ban certain tweets containing images of prisoners of war, that also apply to governments engaged in armed conflict restricting access to the open internet. Twitter aims to strike a balance between supporting essential documentation of events during the conflict and preventing the exploitation by state actors. State-run accounts will no longer be recommended on Twitter's home timeline, explore page, or search results if they originate from countries restricting internet access during armed conflicts.

During the Russian invasion, many researchers sharing primary material from SNs open-source intelligence (OSINT) [2] have faced unexpected account suspensions. Twitter later clarified that these suspensions were unintentional and not part of a coordinated campaign. The platform reinstated access to affected accounts promptly. Twitter's synthetic and manipulated media policy, which addresses the sharing of misinformation, was cited as the possible reason for the suspensions. However, it remains unclear how the violation of this policy was deemed.

Overall, our study contributes to understanding the phe-

nomenon of Twitter suspensions by identifying spam, advertisement campaigns, and scam activities related to the Russia-Ukraine conflict. We provide insights into the characteristics of suspended accounts and analyze the content shared by these accounts. By employing ML models and explainability techniques, we shed light on the factors influencing the decision to suspend user accounts on Twitter.

Implementation source code and the utilized features are available at https://github.com/alexdrk14/TwitterSuspension repository. The dataset is shared in the tabular format containing only the extracted feature values, without any personal user or tweet identifiers for privacy protection.

II. RELATED WORK

Similar works focus on the suspension factors based on Twitter policy topics [3], [4], including the violation of Twitter rules and policies [5], hateful and abusive activities [6]–[8], violation of civil integrity [9] and spam campaigns [10]–[15]. One of the initial and highly influential works was from Vern Paxson [12], examining a dataset of 1.8B tweets. They characterize the spamming methodologies of 1.1M suspended accounts posting 80M tweets, analyze the features, evaluate the abuse of URLs, and provide an in-depth analysis of five spam campaigns. This study first reveals and analyses the marketplace of *spam-as-a-service* on Twitter. Also, in [16] the authors detect compromised accounts in a dataset of 14M victims on Twitter and apply clustering for the suspended and deleted users.

Similar to our study, in [4] after they obtain a Russian, Spanish, and English Twitter corpus on the Russia-Ukraine conflict during 2014, they analyze suspension and deletion factors. The authors utilize multiple approaches of feature extraction based on n-grams, Latent Semantic Analysis, text embeddings, topics, emotions, and images and reveal the major differences between the deleted and suspended accounts. In [17] in 2.7K suspended Twitter accounts they study how dedicated accounts ('trolls') are involved in spreading disinformation, the type of content they disseminate, and their influence on the information ecosystem. In [18] they explore Russian trolls and accounts acting on behalf of the Russian state, in a dataset of 170K control Twitter accounts, highlighting interesting quantitative covariates among the flagged/suspended accounts as an indication of Russian trolls.

An additional work analyzing the Twitter suspension mechanism was done during the 2020 US Presidential Election period [5]. This work shows that suspended accounts have a higher probability of posting hateful tweets, using hashtags related to conspiracy theories belonging mostly to younger accounts, in terms of account registration date.

Additionally, explainability models in ML have also been incorporated in Twitter research towards prediction and sentiment analysis [19], [20], towards identification of bots [14], [21], or fake news detection [22], [23].

Recent studies concentrate on explainable ML implementations, where the models are utilized to describe the differences between suspended and normal accounts. For example, in [24],

Table I: Set of hashtags used in our data collection query in Russian, Ukrainian, and English. Translation for non-English HTs is in parenthesis.

#Ukraine, #Ukraina, #ukraina, #Украина(Ukraine), #Украине(Ukraine), #PrayForUkraine, #UkraineRussie, #StandWithUkraine, #WWIII, #StandWithUkraineNOW, #RussiaUkraineConflict, #worldwar3, #Война(War), #RussiaUkraineCrisis, #StopPutin, #RussiaInvadedUkraine, #BlockPutinWallets, #UkraineRussiaWar, #Putin, #Russia, #Россия(Russia), #StopRussianAggression, #StopRussia, #Ukraine_Russia, #Russian_Ukrainian, #SWIFT, #NATO, #FuckPutin, #solidarityWithUkraine, #PutinWarCriminal, #PutinHitler, #BoycottRussia, #StopWar, #with_russia, #FUCK_NATO, #ЯпротивВойны(ImAgainstWar), #StopNazism, #myfriendPutin, #UnitedAgainstUkraine, #ЯМыРоссия(IWeRussia), #ВеликаяРоссия(GreatRussia), #Путинмойпрезидент(PutinIsMyPresident), #россиявперёд(GreatRussia), #ПутинНашПрезидент(PutinOurPresident), #ЗаПутина(ForPutin), #ПутинВведиВойска(PutinBringTroops), #СЛАВАРОССИИ(GloryToRussia), #СЛАВАВДВ(GloryToAirborneForces)

they analyze the Indian 2019 general elections dataset and compare the differences between three categories of accounts; normal, suspended, and restored. They show the differences between all three classes with SHAP explainability method. This approach reveals that suspended accounts, in comparison with other user categories have a higher retweet rate proportion, average unique hashtags count, and average number of tweets per hour.

Other relevant works on Twitter suspension warnings and patterns include [25], where they demonstrate how suspension warnings on Twitter can influence hate speech. In [26] they study the properties of bullies and aggressors, by extracting text, user, and network-based attributes, as well as the features that distinguish them from regular users. They compare their results with the suspension and deletion of accounts of seemingly undetected users on Twitter. Additionally, authors in [27] explore the patterns of suspended users on Twitter, while they show that removing suspended users has no significant impact on the network structure. Linking Twitter and the suspended accounts, a study in a dataset of 41,352 suspended spammer accounts, [28] reveals the users connected to them, investigates link farming in the Twitter network, and then explores the mechanisms to discourage this activity. Finally, in [29] they show a large Twitter community whose activity supports ISIS propaganda diffusion in varying degrees. They apply Iterative Vertex Clustering and Classification (IVCC) and leverage clustering and Twitter suspensions to infer positive case instances that give partition to the training set with 96% accuracy. In the aforementioned works, they limit their methodology in terms of the extracted characteristics of user profiles, shared content, activity, and user relations. Additionally, most of the studies

Table II: Number of extracted and selected features for each model category.

Feature category	# Total features	# Model selected
Profile	54	36
Activity timing	139	109
Textual	67	53
Post embedding	385	328
Graph embedding	150	140
Combination	1565	197

do not include explainability methods of implemented ML classification models in their analysis. In comparison with the presented approaches, we extract and evaluate a large volume of user features to identify which category is more informative in terms of account suspension. Additionally, we develop ML methods based on each of the feature categories and identify their performance over two different time periods. Finally, we exploit the most accurate models to provide explainability of the model decision with the use of the SHAP game-theoretical approach.

III. Data

For the purposes of our study, we collect a Twitter corpus related to the Russo-Ukrainian War. We use the Twitter streaming API to retrieve public tweets using popular hashtags related to the selected topic (Table I). Our data collection contains 107,735,220 tweets from 9,851,176 users starting from February 23, 2022. To label suspended accounts, we also use Twitter compliance API on a daily basis to identify the exact date of account suspension. Based on this labeling method, we identify 433,466 suspended accounts and 86,630 deactivated accounts. To extract normal users with a high probability of being normal, we remove all suspended/deactivated accounts.

IV. Methodology

As mentioned earlier, in this study we focus on user-shared content, daily activity, and characteristics among the suspended accounts on Twitter, during the 2022 Russia-Ukrainian War. Towards this direction, we extract raw text from tweets of 433,466 suspended accounts with a total volume of 4.8M documents, by applying some primitive filtering (drop URL links and user mentions).

A. Shared content

Initially, we are interested in the shared content across the suspended accounts. Due to the large volume of the extracted documents, the identification of shared content via manual analysis is a challenging task. To overcome this issue we translate posted documents via text embeddings. In our case, we use LaBSE (Language-agnostic BERT Sentence Embedding) [30] model which allows us to overcome two main issues; the multi-language similarity (since our dataset contains posts from a variety of languages) and the normalization of text length and semantics similarity (embeddings of texts with similar semantics are very close to each other). A common

approach to identifying similar topics is to use similarity and clustering methods. However, this methodology is highly time-consuming, especially in our case with large volumes and high vector dimensionality of the extracted embeddings. To handle this issue we apply PCA dimensionality reduction.

This procedure reduces vector dimensionality from 768 to only 20 and reduces the complexity of the text clustering. The developed vectors are clustered via the cosine similarity clustering algorithm which identifies 48,486 unique clusters. We utilize the centroid of each cluster to analyze the text of the shared content.

B. Suspension prediction

Besides the shared content identification, we also investigate whether the feature categories can contribute to a powerful and accurate model for detecting user suspension. This procedure includes the implementation of a model based on each separate feature category, as well as, a combination model that contains all extracted features. Our methodology consists of a model creating a pipeline including feature selection, K-Fold cross-validation, and measurement of the performance on two separated dataset portions.

To develop a generalizable model, we select a restricted time window for the feature extraction and performance estimation. The time restriction helps reduce the model's complexity. In our case, we choose a time window of 21 days, which reduces the time required for user monitoring (in the case of a real-time application) and helps the ML model capture generic patterns.

To identify the features that are effective in a long-term scenario, we extract two separate data portions. The first covers a 21-day monitoring period. This is because we need to simulate a real-case scenario where the initial data are collected for feature selection, model fine-tuning, and initial evaluation. The second portion is used as unseen data for proper model performance evaluation and comparison. Testing the model on the second part of 21 days of data, allows us to measure its performance in a real-case scenario, where the model is fed with data points from different periods of time containing diverse activity patterns. The first dataset includes data from February 23, 2022, until March 15, 2022, and represents the initial training sample. The second portion also contains exactly 21 days, but from a different time period, starting on March 16, 2002, until April 6, 2022. The rest of the paper will refer to these parts as *portion A* and *portion B*.

The extracted data contains 37,195 and 11,716 suspended accounts for *portion A* and *portion B* respectively. Due to the extreme imbalance between normal and suspended accounts, we utilize an under-sampling methodology and select randomly an equal volume of normal users for each data period. Based on the *portion A* data, we extract the following feature categories: user profile, activity timing, textual, post embeddings, and graph embeddings.

Our main goal in this study is to identify which of the presented feature categories are more informative in terms of Twitter account suspension, as well as which feature values have the greatest impact on the model's decision. To achieve

this, we create a model for each feature category and a combination model that contains all the extracted features. The following sections explain the extracted feature categories.

C. Profile features

In this section, we describe the feature extraction of Twitter users' profile objects. The profile objects contain user metadata, including the username, the description, the number of followers and friends, etc. The profile features provide information about the account creator, e.g. how similar the username and screen name are; and whether an account provides a long description, or uses a default profile and background picture. In the case of automatically created accounts, these features may be shared between multiple accounts, while the username and screen name have high similarity. Additionally, it is possible to evaluate the account dynamics according to the profile age, number of followers, number of friends, or number of post activities, using the following formula:

$$by\_age(A) = \frac{A}{Days\ since\ account\ creation} \quad (1)$$

where A is the number of specific actions performed by a user (number of tweets/followers/friends etc.).

In our case, we can also calculate the growth of these parameters, during the monitoring period and calculate the growth of friends, followers, and the number of tweets between the first and the last object of the monitoring feature extraction period. The formula that calculates the growth (G) is:

$$G(A_{start}, A_{end}) = \frac{A_{start} + A_{end}}{A_{start}} \quad (2)$$

where A_{start} is the value of actions from the first user object and A_{end} is the last action value, during the feature extraction period. Based on our *portion A* data, we extract in total 54 unique profile features (Table II).

D. Activity timing features

A particular feature category is based on the statistical measurements of a user's actions (shared posts). We measure two categories of user activity patterns. The actions according to which part of the day they happen (hour) and the index of the day within a week. The second category measure is the statistics of the collected user posts (tweets, retweets, and quotes). In the case of activity per hour of the day, we measure the number of user actions (tweets, retweets, and quotes), according to the daily hours. Based on these measurements, we calculate the percentage of the user's activity and identify the hours within a day that users are most active. This feature provides the unique characteristics of the user activity during the day.

Similarly, we compute the user activity, according to the days of the week. This feature may also provide unique information about the user activity; for example, identify whether the users are more active during the weekends.

Furthermore, we measure the statistics of the user actions during the feature extraction period. For this purpose, we

Table III: Examples of detected discussion topics from suspended accounts. Sensitive metadata like account, mention, hashtag, and URL info is removed for privacy reasons.

Category	Text
Crypto/NFT	Knock knock knock... Anybody is there? Your lucky door knocking ; RETWEET TAG 3 friend 1000$ #Bitcoin #Airdrop #StopWar #UkraineRussia #StopRussia #Crypto #NFTs #NATO #worldwar3 #PrayingForUkraine #Putin #Giveaway #ETH #cryptocurrency
Spam	How to find over 100 ways to earn money with URL via COMPANY #Ukraine #RussianArmy #AssassinsCreed #KingCharlesIII #QueenElizabeth #earthquake #USOpen #QueenElizabethII
Injection	#Ukraine needs weapons and humanitarian assistance to defend against #Putin. Russian troops shoot a nuclear power plant. Stop innocent civilian deaths. People around the world ask NATO to close the airspace over Ukraine. MENTION, provide #SafeAirliftUkraine

utilize a certain statistical measurement based on the action time (retweet time and quote time), where we calculate the time difference between the original post creation and the user response action time. This metric shows the average time between the user's reaction to other users' activity. The accounts that have a very "quick" reaction time are probably automated accounts that are controlled by computer software.

Additionally, we identify the user tweets, retweets, and quotes that are the most common forms of user activity, by measuring their percentage. These described measurements provide information about the user activity routine, like the active hours of the day, the days of the week, and the general user profile preferences of content sharing.

E. Textual features

Except from the information that can be extracted from the user profile and action pattern, we need to collect the content that users share, since Twitter users mostly communicate via

Table IV: The most popular discussion topics according to our manual analysis, combined with the total amount of analyzed tweets and identified clusters.

Topic	# Tweets	# Clusters
Giveaway/Crypto/NFT	165,226	18
Content injection	71,931	8
SPAM/advertisment	64,699	9
Military Intelligence	59,367	15
Support for Ukraine	34,952	2
Total analyzed	4,808,904	48,486

content sharing. For this reason, we utilize two techniques of content analysis. Initially, we extract content metadata statistics, regarding the semantics users adopt to enrich their shared content and the text semantics via sentence embedding. In this section, we describe the first category based on the text metadata semantics. We separate this feature category into three subcategories: tweets, retweets, and quotes. For each one, we extract separate metadata since they are created under different contexts (e.g. RTs are not written by the user herself, but already created by another user).

For each category, we retrieve the number of hashtags, URLs, and mentions. For each of these measurements, we take into account the minimum, maximum, average, and standard deviation. These metrics not only provide a picture of the volume but also describe the distribution of the particular feature usage. Also, since some of the hashtags are very popular, while others may be uniquely created and used only by a particular user, we compute the Term Frequency - Inverse Document Frequency (TF-IDF). The TF-IDF shows whether a particular hashtag is trending across other users or not. This information allows us to identify whether an account utilizes mostly popular entities (hashtags, user mentions). The last measurement of content metadata we compute is the user vocabulary size.

F. Post embedding features

In the previous section, we describe how we extract metadata from the user-shared objects. However, this analysis lacks text semantics. Our collected dataset contains posts in 65 different languages, which makes the extraction of text semantics very challenging. In order to eliminate this issue, we use pre-trained text embedding solutions. We utilize the LaBSE (Language-agnostic BERT Sentence Embedding) Neural Network model [30] which is trained over more than 100 languages. The particular model was selected for two reasons; it provides a large space of already trained languages and does not require the alignment of input text sequences.

Based on the selected model, we parse all user posts and export 384 features for each of them. Additionally, we store the categories of the user posts (tweets, retweets, or quotes) as one hot encoding. This particular information allows the transformation of textual information (that can't be processed by an ML model) into a numeric vector space and allows the processing by a classification model.

G. Graph embedding features

Twitter as a modern SN provides multiple social interactions between registered users (retweets, mentions, and quotes). In this section, we answer the research question of "Which type of social relations are important for suspension detection?". To answer this question we extract the social relations between the users in a graph form. In this social graph, the nodes represent the users, and the edges represent the relations between them.

Graph representations cannot be used directly by an ML model for training and prediction. In order to solve this issue

Table V: Graph embedding evaluation performance over different social relation graphs combined with different output dimensions. The highest performance is presented in bold text.

Social Graph	Output dimension	MRR	AUC
Quote	50	0.65	0.81
Mention	50	0.54	0.87
Retweet	50	0.54	0.84
Multi-layer	50	0.79	0.96
Quote	100	0.67	0.78
Mention	100	0.53	0.87
Retweet	100	0.52	0.84
Multi-layer	100	0.82	**0.97**
Quote	150	0.66	0.78
Mention	150	0.52	0.86
Retweet	150	0.52	0.84
Multi-layer	150	**0.84**	**0.97**
Quote	200	0.66	0.77
Mention	200	0.52	0.85
Retweet	200	0.52	0.84
Multi-layer	200	0.54	0.80

modern solutions utilize Neural Network (NN) implementations [31]. In the case of a NN implementation, the model learns the graph relations between the nodes and provides a representation (also known as node embedding) in the form of a numerical vector for each relational node of the graph.

The provided representation allows the utilization of the graph relational knowledge as an input to the ML model. For this purpose we replicate PyTorch Big Graph Neural Network model [32], since the particular implementation allows us to feed and process very large graphs with millions of nodes, without any restriction concerning the number of social relations (also known as graph layers). The particular implementation allows us to experiment between the single relation and the multi-relational graph embeddings and keep the original structure and parameters of the GNN [32].

Twitter provides multiple user-to-user interactions, namely retweets, mentions, and quotes. In our implementation, we are interested in the identification of social interactions that provide the most important information about a user-to-user relationship. For this purpose, we extract four different social interaction graphs, quotes, mentions, retweets, and a combination of them (also known as a multi-layer graph). As a metric of social interaction importance, we keep learning the performance (MRR and AUC) of our Neural Network model for each graph, since the performance of the NN model affects immediately the performance of the ML classification model.

We permit our model to learn accurately the user-to-user relations since it achieves higher MRR and AUC scores according to the presented performance in Table V. Specifically, the multi-relational graph as a combination of social interactions has an embedding size equal to 150.

V. Experimental Results

A. Shared content

In this chapter, we present the results, based on the experiments described in the previous section. Specifically, we identify that the most popular category of suspended content is correlated to 'Crypto', 'NFT', and 'giveaways'. We assume that content creators in order to trigger user attention, take advantage of trending topics, and keywords referring to the Russo-Ukrainian War, especially during the first months of the war.

Additionally, we identify similar content injection, where the users share identical content and mention celebrities (Table III) to gain popularity. This kind of activity pattern is very similar to botnet activity where a particular group of accounts shares identical content at almost identical times and takes advantage of popular topics and users, as shown in similar studies [33]. Besides similar content, we also identify large spam and advertisement campaigns where accounts maliciously ride on the success of popular hashtags and trending topics to promote their products [13].

Since we identify that the most popular discussion topics were 'Crypto' and 'NFT', we initialize a search for related keywords like 'crypto', 'NFT', and 'donation' within the clusters. It seems that the keywords of 'crypto' and 'donation' are very popular across the suspended users' conversations [34]. After a dynamic search and manual inspection, we reveal that multiple accounts are sharing messages similar to:'*Stand with the people of Ukraine. Now accepting Bitcoin donations: BTC*' (Table III). These messages trigger the attention of users to donate money in the form of Bitcoin and Ethereum with a wallet address. More specifically, we count 14 Bitcoin and 9 Ethereum wallets with a total of 16 transactions since the Russian invasion and collect in total 0.02896581 Bitcoins and 0.97419 Ethereum coins. More specifically, we expose these wallets posting tweets that support both sides, Ukraine and Russia. In Table III we show some examples of tweets in an effort to steal non-fungible tokens (NFTs) (usernames and mentions that have been removed).

Furthermore, we execute measurements of content toxicity posted by suspended accounts, in order to identify whether toxicity plays an important role in the current suspension decisions. To answer this question we use a Neural Network model [35], already trained over multiple languages. Our analysis shows that only 2.1% of the suspended account posts are toxic. Based on this result we assume that toxicity did not play an important role in the Twitter suspension decision.

B. Performance

Besides the content shared of suspended accounts, we also focus on building an ML model and a feature set that can achieve high detection performance. To evaluate the performance of our model, we utilize multiple testing sets with K-Fold cross-validation. During the initial step, we measure the performance of our models based on the average F1-score, during the K-Fold cross-validation (K=5). The cross-validation measurement provides the first impression of the model performance over the training and validation of the data set portion. In the following step, we measure the performance of each model, based on the testing (hold-out) data not seeded during the feature selection, the model fine-tuning, and the cross-validation. The performance of the test data provides crucial information about the model generalization ability and the over/under-fitting issues. Based on these results (see Table VI) of the K-Fold cross-validation and test data-set performance, we identify that the best performance is achieved by a model that contains the combination of the features and the user profile model, with a slightly better performance of the combination model over the validation and test data-sets.

Since both validation and test data portions were extracted from the exact same time period, we are interested in the identification of the performance of new data created by the users in different time periods. For this reason, we retrieve a second testing dataset (*portion B*) containing unseen normal and suspended users from the next 21 days. The extracted dataset contains exactly the same feature categories with *portion A*. Each feature is extracted based only on the *portion B* 21 days period, in order to avoid information leaking, except the graph embedding. In the case of graph embeddings, we keep the training data user relation as well, in order to properly allocate the node positions in an N-dimensional output space.

The extracted dataset contains exactly the same feature space and keeps class balance within the dataset, with the usage of an under-sampling method, as well as within the training, validation, and test data portions. During the performance measurement, we train our model over the entire *portion A* dataset and measure how the model performs on unknown data originating from a different time period (*portion B* data). Based on the results presented in Table VI, we notice that the performance of multiple models drops significantly, especially in the case of embeddings (text and graph). The drop in the embedding features indicates that they are strongly correlated to the user content (or activity, in the case of graph embeddings). Additionally, we notice that a slight change in the user actions leads to completely different results of the embedding extraction technique.

Despite this, some models achieve decent performance, even in the next time period, such as the profile and the combination model. The profile model provides the lowest performance drop, in comparison with any other model. This outcome allows us to assume that the user profiles remain almost the same even after a 21-day period.

C. Model explainability

Based on the measured performance, seen in Table VI, we manage to identify two models that allow us to detect suspended accounts. Additionally, within this study, we are interested in the identification of differences between normal and suspended accounts on Twitter. Toward this goal, we build an ML model to spot the differences between the feature values of suspended and normal accounts. We achieve this with the usage of state-of-the-art SHAP (SHapley Additive exPlanations) [1] method, based on a game theoretic approach

Table VI: The model performance is measured during the K-Fold Cross Validation (avg. score). We show the test set of the initial *portion A* data, and the second test containing *portion B* data. Each model is trained/tested over the described set of features.

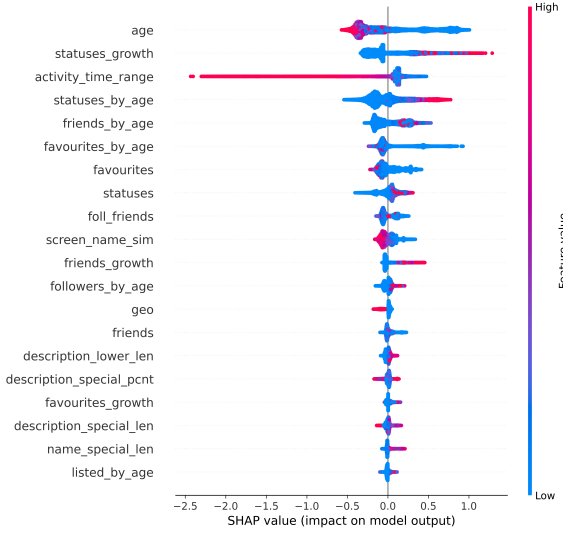| Model	Validation			Test			Second Test		
	F1	ROC-AUC	Acc.	F1	ROC-AUC	Acc.	F1	ROC-AUC	Acc.
Profile	<u>0.86</u>	<u>0.93</u>	<u>0.86</u>	<u>0.86</u>	<u>0.94</u>	<u>0.87</u>	**0.79**	**0.90**	**0.81**
Activity timing	0.67	0.76	0.70	0.68	0.77	0.71	0.45	0.62	0.58
Textual	0.71	0.82	0.75	0.72	0.82	0.75	0.44	0.63	0.59
Post embedding	0.85	0.92	0.86	0.83	0.91	0.84	0.00	0.73	0.50
Graph embedding	0.77	0.86	0.79	0.77	0.87	0.79	0.21	0.50	0.50
Combination	**0.88**	**0.95**	**0.88**	**0.88**	**0.95**	**0.88**	<u>0.75</u>	<u>0.89</u>	<u>0.79</u>

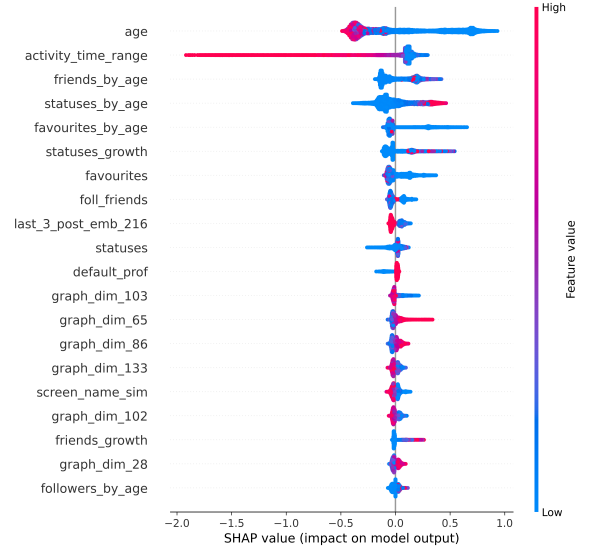Figure 1: Explainability figure (SHAP) of profile feature trained models.

Figure 2: Explainability figure (SHAP) of combination feature trained models.

to explain the output of the ML model, and considered one of the best models in interpretability [36]. We select the profiles and the combination of the feature models, in order to explain these value differences.

In Fig. 2 and 1 we can see that both models identify similar patterns; suspended accounts have a low account age and a short period of activity (specifically notice the 'activity_time_range' in Fig. 2 and 1) and at the same time high values of status, according to the account age. These results show that suspended accounts are mostly fresh registered accounts posting high volumes of tweets, within a short time period. Additionally, suspended accounts grow their friends' relationship networks faster than normal users. Based on the identified results, the activity patterns of the suspended accounts on Twitter are very similar to those of the automated bot accounts.

VI. Conclusion

In this study, we extensively examine and extract a vast amount of characteristics for the suspended accounts involved in the online discourse during the Russo-Ukrainian War,

initiated in 2022. For this purpose, we collect a large volume of user profiles along with their shared content. Based on the examined features we develop and evaluate multiple ML models and identify their user characteristics. Our methodology allows the accurate prediction of the suspended accounts by Twitter. Specifically, we utilize a state-of-the-art model decision explainability SHAP methodology in order to describe the actual model decisions based on the feature values. Our study shows that the most important characteristics behind suspension are the age of the account combined with the user actions (posts, likes, retweets, number of friends, and followers). Furthermore, we show that the most informative feature category is the combination of multiple features with 0.88 F1 and 0.95 ROC-AUC scores during the initial *portion A* data. In the case of the *portion B* data, we estimate the performance prediction on long-term scenarios, where profile features achieve the highest performance of 0.79 F1 and 0.9 ROC-AUC. Based on the achieved performance we are interested in further analysis over the larger time periods. Such an approach will allow us to evaluate and compare all of the

models based on the described feature categories.

VII. Acknowledgement

References

[1] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," arXiv preprint arXiv:1705.07874, 2017.

[2] T. Verge, "Twitter accounts sharing video from ukraine are being suspended when they're needed most." https://shorturl.at/jALOU, 2022, accessed: 2023-07-19.

[3] F. A. Chowdhury, L. Allen, M. Yousuf, and A. Mueen, "On twitter purge: A retrospective analysis of suspended users," in Companion Proceedings of the Web Conference 2020, 2020, pp. 371–378.

[4] S. Volkova and E. Bell, "Identifying effective signals to predict deleted and suspended accounts on twitter across languages," in Proceedings of the International AAAI Conference on Web and Social Media, vol. 11, no. 1, 2017, pp. 290–298.

[5] F. A. Chowdhury, D. Saha, M. R. Hasan, K. Saha, and A. Mueen, "Examining factors associated with twitter account suspension following the 2020 us presidential election," in Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2021, pp. 607–612.

[6] A. M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis, "Large scale crowdsourcing and characterization of twitter abusive behavior," in Twelfth International AAAI Conference on Web and Social Media, 2018.

[7] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in Proceedings of the International AAAI Conference on Web and Social Media, vol. 11, no. 1, 2017, pp. 512–515.

[8] P. Nakov, V. Nayak, K. Dent, A. Bhatawdekar, S. M. Sarwar, M. Hardalov, Y. Dinkov, D. Zlatkova, G. Bouchard, and I. Augenstein, "Detecting abusive language on online platforms: A critical analysis," arXiv preprint arXiv:2103.00153, 2021.

[9] E. Ferrara, H. Chang, E. Chen, G. Muric, and J. Patel, "Characterizing social media manipulation in the 2020 us presidential election," First Monday, 2020.

[10] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Detecting automation of twitter accounts: Are you a human, bot, or cyborg?" IEEE Transactions on dependable and secure computing, vol. 9, no. 6, pp. 811–824, 2012.

[11] X. Zhang, S. Zhu, and W. Liang, "Detecting spam and promoting campaigns in the twitter social network," in 2012 IEEE 12th international conference on data mining. IEEE, 2012, pp. 1194–1199.

[12] K. Thomas, C. Grier, D. Song, and V. Paxson, "Suspended accounts in retrospect: an analysis of twitter spam," in Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference, 2011, pp. 243–258.

[13] D. Antonakaki, I. Polakis, E. Athanasopoulos, S. Ioannidis, and P. Fragopoulou, "Exploiting abused trending topics to identify spam campaigns in twitter," Social Network Analysis and Mining, vol. 6, no. 1, pp. 1–11, 2016.

[14] A. Shevtsov, C. Tzagkarakis, D. Antonakaki, and S. Ioannidis, "Identification of twitter bots based on an explainable machine learning framework: The us 2020 elections case study," Proceedings of the International AAAI Conference on Web and Social Media, vol. 16, no. 1, pp. 956–967, May 2022. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/19349

[15] D. Antonakaki, P. Fragopoulou, and S. Ioannidis, "A survey of twitter research: Data model, graph structure, sentiment analysis and attacks," Expert Systems with Applications, vol. 164, p. 114006, 2021.

[16] K. Thomas, F. Li, C. Grier, and V. Paxson, "Consequences of connectivity: Characterizing account hijacking on twitter," in Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, 2014, pp. 489–500.

[17] 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 Companion proceedings of the 2019 world wide web conference, 2019, pp. 218–226.

[18] 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," in 12th ACM Conference on Web Science, 2020, pp. 1–10.

[19] K. Fiok, W. Karwowski, E. Gutierrez, and M. Wilamowski, "Analysis of sentiment in tweets addressed to a single domain-specific twitter account: Comparison of model performance and explainability of predictions," Expert Systems with Applications, vol. 186, p. 115771, 2021.

[20] F. Morstatter, L. Wu, T. H. Nazer, K. M. Carley, and H. Liu, "A new approach to bot detection: striking the balance between precision and recall," in 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, 2016, pp. 533–540.

[21] M. Kouvela, I. Dimitriadis, and A. Vakali, "Bot-detective: An explainable twitter bot detection service with crowdsourcing functionalities," in Proceedings of the 12th International Conference on Management of Digital EcoSystems, 2020, pp. 55–63.

[22] J. C. Reis, A. Correia, F. Murai, A. Veloso, and F. Benevenuto, "Explainable machine learning for fake news detection," in Proceedings of the 10th ACM conference on web science, 2019, pp. 17–26.

[23] E. Puraivan, E. Godoy, F. Riquelme, and R. Salas, "Fake news detection on twitter using a data mining framework based on explainable machine learning techniques," 2021.

[24] A. Kapoor, R. R. Jain, A. Prabhu, T. Karandikar, and P. Kumaraguru, ""i'll be back": Examining restored accounts on twitter," in IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2021, pp. 71–78.

[25] M. M. Yildirim, J. Nagler, R. Bonneau, and J. A. Tucker, "Short of suspension: How suspension warnings can reduce hate speech on twitter," Perspectives on Politics, pp. 1–13, 2021.

[26] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, "Mean birds: Detecting aggression and bullying on twitter," in Proceedings of the 2017 ACM on web science conference, 2017, pp. 13–22.

[27] W. Wei, K. Joseph, H. Liu, and K. M. Carley, "Exploring characteristics of suspended users and network stability on twitter," Social network analysis and mining, vol. 6, no. 1, pp. 1–18, 2016.

[28] S. Ghosh, B. Viswanath, F. Kooti, N. K. Sharma, G. Korlam, F. Benevenuto, N. Ganguly, and K. P. Gummadi, "Understanding and combating link farming in the twitter social network," in Proceedings of the 21st international conference on World Wide Web, 2012, pp. 61–70.

[29] M. C. Benigni, K. Joseph, and K. M. Carley, "Online extremism and the communities that sustain it: Detecting the isis supporting community on twitter," PloS one, vol. 12, no. 12, p. e0181405, 2017.

[30] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic bert sentence embedding," arXiv preprint arXiv:2007.01852, 2020.

[31] S. W. Min, K. Wu, S. Huang, M. Hidayetoğlu, J. Xiong, E. Ebrahimi, D. Chen, and W.-m. Hwu, "Pytorch-direct: Enabling gpu centric data access for very large graph neural network training with irregular accesses," arXiv preprint arXiv:2101.07956, 2021.

[32] A. Lerer, L. Wu, J. Shen, T. Lacroix, L. Wehrstedt, A. Bose, and A. Peysakhovich, "Pytorch-biggraph: A large scale graph embedding system," Proceedings of Machine Learning and Systems, vol. 1, pp. 120–131, 2019.

[33] A. Shevtsov, M. Oikonomidou, D. Antonakaki, P. Pratikakis, A. Kanterakis, P. Fragopoulou, and S. Ioannidis, "Discovery and classification of twitter bots," SN Computer Science, vol. 3, no. 3, pp. 1–29, 2022.

[34] DomainToolsResearch, "Crypto winter: Fraudsters impersonate ukraine's government to steal nfts and cryptocurrency," https://www.domaintools.com/resources/blog/fraudsters-impersonate-ukraines-government-stealing-nfts-crypto/, 2022.

[35] L. Hanu and Unitary team, "Detoxify," Github. https://github.com/unitaryai/detoxify, 2020.

[36] E. Mosca, F. Szigeti, S. Tragianni, D. Gallagher, and G. Groh, "Shap-based explanation methods: a review for nlp interpretability," in Proceedings of the 29th International Conference on Computational Linguistics, 2022, pp. 4593–4603.