IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 11, 2021, accepted July 26, 2021, date of publication August 5, 2021, date of current version August 16, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3102865

Analyzing the Existence of Organization Specific Languages on Twitter

RUBÉN SÁNCHEZ-CORCUERA[1], ARKAITZ ZUBIAGA[2], AND AITOR ALMEIDA[1]
[1]DeustoTech, University of Deusto, 48007 Bilbao, Spain
[2]Social Data Science Laboratory, Queen Mary University of London, London E1 4NS, U.K.

Corresponding author: Rubén Sánchez-Corcuera (ruben.sanchez@deusto.es)

This work was supported in part by the Basque Governments Department of Education, and in part by the Deustek Research Group under Grant IT1078-16D.

ABSTRACT The presence of organisations in Online Social Networks (OSNs) has motivated malicious users to look for attack vectors, which are then used to increase the possibility of carrying out successful attacks and obtaining either private information or access to the organisation. This article hypothesised that organisations have specific languages that their members use in OSNs, which malicious users could potentially use to carry out an impersonation attack. To prove these specific languages, we propose two tasks: classifying tweets in isolation by their author's organisation and classifying users' entire timelines by organisation. To accomplish both tasks, we generate a dataset of over 15 million tweets of five organisations, and we apply language dependant models to test our hypothesis. Our results and the ablation study conclude that it is possible to classify tweets and users by organisation with more than three times the performance achieved by a traditional ML algorithm, showing a substantial potential for predicting the linguistic style of tweets.

INDEX TERMS Social network services, Twitter, transformers, natural language processing, attack vector.

I. INTRODUCTION

Online Social Networks have become vital for many people and organisations (companies, NGOs or political parties, for example), providing them with new communication channels. Thanks to these new channels, an organisation can communicate with their members or clients, update them on the latest developments and decisions, or use these platforms for marketing purposes. Unfortunately, the increasing popularity of these platforms has also attracted the attention of malicious users who, from the beginning, have been committing attacks against users, organisations, or the social networks at large [1].

Even though the attacks are being carried out on many social networks and using different methods, the microblogging social network Twitter prevails as a platform that receives a substantial amount as evidenced by researchers [2]–[4], and the social network itself [5]. Furthermore, a body of research in this direction has highlighted the presence of such attacks, for example, by producing misinformation during the 2016 US election [6]–[8], idea induction attacks during the Brexit campaign in the UK [9], [10] or attacks between different groups during the Catalan referendum on October 1st, 2017 [11], among others.

As posited in this research, these attacks were in part possible thanks to malicious users' involvement, usually referred to as content polluters or trolls. These users are dedicated to coordinate malicious actions against different ideas, organisations or people within social networks. Malicious users employ different techniques and strategies to amplify their messages and actions, thus reaching more people [12], [13].

One of the strategies that malicious users employ to conduct influential and credible attacks against organisations is infiltrating them by creating fake users with compelling features. This technique aims to create trust bonds and links with members of the attacked organisation by leveraging humans' tendency to build homophilic connections [14]. Homophily refers to humans' tendency to relate and bond with those similar to them according to characteristics such as personality, behaviour or taste; homophily can apply to a wide range of dimensions, including gender, age, organisational role or class. Therefore, malicious users could employ data extracted from users profiles to create homophilic links with them.

For creating fake avatars, malicious users employ profile pictures of real people or images generated by Generative Adversarial Networks (GANs) [15], [16], capable of

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

generating reliable faces. They also use legit information in their biography and post tweets related to the topics discussed in the group they want to attack [4], [17]. Because of the use of real information and topics related to the groups they want to attack, malicious users can establish trust bonds and links that can ease the attacks [1], [18].

We hypothesise that members of organisations use a specific language that malicious users may copy to impersonate or create trust bonds to attack them on Twitter. These types of impersonation attacks are common on social networks such as Twitter and Instagram. On Twitter, the most famous case of impersonation is that of Warren Buffett,[1] in which an account was created imitating his language and using his image and biography. On Instagram, on the other hand, impersonations are done by using photos of the target person to create an account with the same characteristics as the original ones and comment on posts to gain visibility and spread polluted content [19].

Therefore, motivated by the latest attacks and works done around security on social networks, we question whether the common language among the organisation members can present an attack vector for the organisation as a whole. To this end, we proposed an analysis to determine if organisations and their members use specific and distinguishable languages to detect and prevent attacks against them using these techniques.

A. OBJECTIVES AND CONTRIBUTIONS

To the best of our knowledge, this is the first study of social media language to predict the organisation to which a Twitter user belongs. We hypothesise that it is possible to know from which organisation a user is by merely using the language of their tweets. This hypothesis, in turn, enables tackling our research question, which states that organisations have specific languages.

To conduct the study, we divided the analysis into two different tasks: (1) classification of tweets in isolation to predict the organisation its author belongs to (tweet-level), and (2) classification of users by organisation by using their entire timeline of tweets (user-level). First, we aim to demonstrate that individual members consistently use a specific language that links it to their organisations with the tweet-level task. Second, with the user-level task, we want to demonstrate that the specific language can also be inferred when using all of a user's tweets in aggregation. Based on this, we set forth two key research questions that we tackle in this work:

- **RQ1.** Can the organisation of a social media user account be determined by the aggregated content of their posts?
- **RQ2.** Is each of the individual posts of a social media user account indicative of the organisation to which they belong?

[1]https://www.darkreading.com/analytics/anatomy-of-a-social-media-attack/a/d-id/1326680

To tackle these two research questions, we create a dataset composed of members of five different organisations. For each of these organisations, we identify user accounts for members with social media presence, and we retrieve Twitter profiles and timelines of tweets for each of them. We address the problem as a classification task, using two models that leverage linguistic features of tweets to classify them and compare their performances against a traditional ML algorithm. The prediction performance of the employed models needs to be above an established baseline to confirm RQ1 and RQ2. We complement our study by conducting an ablation study to analyse the different elements that a tweet may have (hashtags, mentions and URLs) and how they contribute to the classification.

The main contributions of the paper can be summarised as:

- The first study that classifies the specific language used by members in different organisations on Twitter.
- An ablation study conducted on the elements of tweets to determine the importance of each one in identifying the belonging of the users to organisations.
- Classification of users and tweets with over 75 points in F1-score, supporting the idea that organisations have differentiated languages.
- A dataset with more than 15 million tweets and more than 25,000 users of 5 different organisations.

II. RELATED WORK

Twitter users have been targets of attacks carried out by malicious users that aim to gather private information [20]–[22] or the promotion of their ideas [23], [24]. Attacks on Twitter are usually conducted by avatars created by malicious users that mimic human behaviour to impersonate real people. Researchers in this area focused on creating frameworks that, using features from Twitter users, can identify if they are malicious users [25]–[28]. The authors of these works use tweets to analyse the language of and characterise malicious users on Twitter, which differs from our objectives of analysing the language of potential targets. Language has also been used to detect the promotion of misinformation or idea induction campaigns on Twitter [23]. Authors of [29] also discovered that malicious users change the style of their posts, therefore, the language, to achieve different objectives. These works show that language is an attack vector that malicious users may use to attack users on ONSs.

A related line of research to ours is that looking at the linguistic style-based text categorisation. In this task, researchers try to classify texts according to their writing style; similarly, we try to classify tweets and users in different organisations. As a first approach, authors of [30] propose the classification of articles from 4 different newspapers and magazines to demonstrate that the style of the text was independent of the article's topic. Later works proposed the categorisation of text by author and genre using intrinsic features from the analysed texts [31], [32]. It has also been tried to combine corpus extracted from different

platforms, social networks and movie review sites in this case, to show that languages from different platforms can help classify the authors of the posts more effectively in some cases [33]. Finally, authorship identification has also been employed in OSNs to help forensics in cases that require the authorship identification for some post [34]. Derived from these tasks, the PAN workshop[2] focused on shared tasks on digital text forensics and stylometry as an annual challenge on authorship identification. Participants of this shared task are invited to develop models to classify texts from authors in 5 different languages using lexical, syntactical, structural and content-specific features [35]. These works demonstrate that the author's style remains even when the topic of the text changes. Whereas they are not focused on determining group or organisation-specific language styles, they provide encouraging evidence to support our hypothesis that language differs across organisations.

A body of research has focused on generative adversarial networks [36]; these can generate text from a corpus and thus mimic tweets from different groups adapting to different styles. An intersection between this technique and style classification is given in the style transferring task [37]. This idea aims to transfer the style from a writer or a text to another. Works by Jhamtani *et al.* [38], and Carlson *et al.* [39] propose different models to transfer Shakespeare style or different bible versions styles to other texts, respectively. Regarding our article, this idea may be used to adapt tweets written by malicious users using the style from the organisation, helping the attacker to seem more like a legit member. We believe that conducting this analysis will make organisations aware of the attack vector that their specific language may represent.

III. DATASET

For our experiments, we selected five organisations from different fields with many employees active on social media. These organisations are:

- Organisation A is an NGO focused on humans' rights.
- Organisation B is a multinational aerospace corporation.
- Organisation C is a multinational professional services network.
- Organisation D is a political party.
- Organisation E is a multinational technology company.

As we can see, some of the organisations belong to different fields, whereas others are similar, such as the technological ones. We have selected these organisations to create a complete dataset that includes both the differences between sectors and the similarities between organisations.

A. DATA GATHERING

For each of the organisations under study, we collected data from the Twitter accounts of their members. We used an algorithm proposed in [40]. It was initially developed for data collection from Facebook, so we adapted it to match the characteristics of Twitter. This algorithm (Algorithm 1) requires a seed list of manually identified users $U = u_1, \ldots, u_n$ to start with, which have been identified as of the organisation at hand. Then, for each seed user u_i, it retrieves their connections, i.e. followers and followees and checks if they belong to the organisation by looking for a predefined set of keywords on their Twitter biography, treating them as members of the organisation if one keyword is found. Subsequently, the algorithm continues with this process with the following user in the queue, sorted using a priority degree assigned by the times a user appears as a follower or followee of the previously analysed users. Whereas the need for the organisation name or a keyword to appear in the biography of users may be a feature with limited availability and limit the generalisability of the approach, the fact that we rely on organisations with large numbers of members reduces the impact of this issue.

For each organisation, we introduced 20 manually identified seed users and ran the algorithm for two weeks. The number of users and tweets collected for each organisation can be seen in Table 1.

Algorithm 1 Organisational Mining Algorithm From [40] Modified for Twitter. The Default Priority Is Set in 30

Input: A set of seed Twitter Usernames (S) of organisation's employees and a set of words related to the target organisation, N

Output: A set of Twitter profiles with twitterid, name, username, biography, followers and followings

Organisational Miner():

$Q \leftarrow Priority - Queue()$
$\forall_{Username \epsilon S}, Q.Enqueue(Username : 30)$
$Crawled \leftarrow \emptyset$
$NonRelatedUsers \leftarrow 0$
while $(Q \neq \emptyset \ \& \ max(Q.Priority) \neq 1 \ \& \ NonRelatedUsers < 1000)$ **do**
  $Username \leftarrow Q.Dequeue()$
  $Page \leftarrow DownloadTwitterProfileData(Username)$
  $Crawled \leftarrow Crawled.append(Username)$
  if *N in Page* **then**
    $Connections \leftarrow ExtractConnFromTwitter()$
    $Connections \leftarrow Connections - Crawled$
    for $(Connection \epsilon (Connections \cap Q))$ **do**
      $Increasepriority(Connection)$
    end for
    for $(Connection \epsilon (Connections - Q))$ **do**
      $Q.Enqueue(Connection, Priority : 1)$
    end for
    $CollectedPages.append(Page)$
  else
    $NonRelatedUsers++$
  end if
end while
return *CollectedPages*

[2]https://pan.webis.de/

TABLE 1. Number of users and tweets per organisations. The last column on the right indicates the number of tweets left per organisation after applying the transformations explained in the last paragraph of Section III to the tweets.

	Users	Tweets	Tweets preprocessed
Organisation A	2,298	1,992,434	885,975
Organisation B	684	620,836	390,985
Organisation C	3,591	2,256,832	1,384,715
Organisation D	7,374	15,640,422	6,270,414
Organisation E	11,099	14,693,852	6,870,413

After two weeks, we stopped the algorithm and harvested the maximum number of tweets for each user allowed by the Twitter API, restricted to the last 3,200 tweets. The dataset is available in Github.[3]

B. PREPROCESSING

Out of the tweets collected for each user, we filtered English language tweets for consistency. We also removed the emojis from tweets. Many works have improved the performance of text classifying models by changing the emojis present in the text by their contextual meaning [41], [42]. Nevertheless, we decided not to change the emojis for their meaning to avoid modifying the original tweets. Furthermore, some emojis may not mean the same in every context, for example, when using irony, and thus, the automatic change of these emojis may alter the meaning of the original tweet. We also removed retweets and very similar tweets (e.g. differing only in a hashtag or a URL) due to their redundant content. We removed all the mentions of their organisations for the remaining tweets, intending to avoid direct mentions of the ground truth label. The number of resulting tweets per organisation can be seen in Table 1.

IV. RESEARCH METHODOLOGY

We propose the gathering and classification of tweets written by members of the selected organisations (RQ2) and to predict users' organisations by aggregating their tweets (RQ1) by using a fine-tuned BERT model [43] and a CNN-based model. We compared the results obtained by these two models against a Random Forest classifier as a baseline against language models. Obtaining good results on both tweet and user tasks will confirm specific organisational languages on Twitter. As tweets may contain elements that could provide external information to classifiers, such as hashtags, mentions or URLs, we also conducted an ablation study removing those components and evaluating their impact on the classification.

A. MODELS

We employed two models that rely on language features (BERT and Multi-CNN) and a Random Forest classifier as the baseline to conduct the text classification task. The BERT-base model contains an encoder with 12 Transformer blocks, 12 self-attention heads, and a size of 768 for the hidden layers. Next, BERT takes an input of a sequence of no more than 512 tokens and outputs the representation of that sequence. Finally, BERT adds a final hidden layer with the same size as the number of possible classification labels and a softmax function for text classification tasks.

BERT is already pretrained for text classification; however, fine-tuning needs to be conducted to adapt the model to a specific task. Therefore, we followed a strategy proposed in [44] consisting of unfreezing several hidden layers and training the model before running the tests. We tested unfreezing several layers starting from the end to obtain the best results for this task. Finally, we selected the last three layers and the classification layer for training as they obtained the best results in the test. For the implementation of BERT, we used the PyTorch version of the Transformer library developed by HuggingFaces available on GitHub[4]

The second language dependant model employed for this task is called Multi-CNN. We opted to use convolutions because of the capacity that they have proved to solve text classification tasks [45], [46] and also those related explicitly to tweet classification [47], [48].

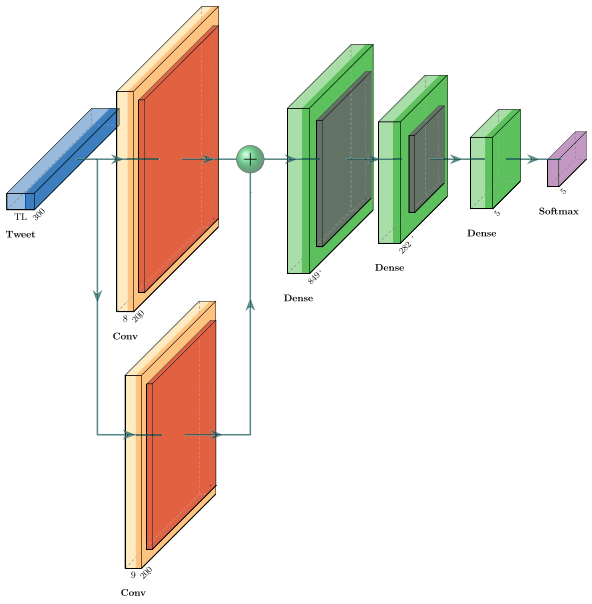

FIGURE 1. Plot of the multi-CNN model.

The model receives the tweets with a fixed size of D × 300 (blue layer in Figure 1), with being D the length of the most extended tweet in the batch and 300 the size of the embeddings. Next, tweets are fed into the convolutional neural networks (orange layers in Figure 1). This network is composed of two different convolutions with region sizes of 8 and 9, respectively. As stated by [49], the use of multiple filters are recommended to learn complementary features about the analysed text. Therefore, each convolution produces

[3]https://github.com/rubensancor/Organisation_language_twitter

[4]Transformers codebase: https://github.com/huggingface/transformers

200 filters to which the activation function is applied, in this case, ReLu. The result of each convolution is a vector of $B \times 200$, being B, the number of tweets introduced to the convolution. Each convolution is followed by a ReLu function and a max-pooling layer to reduce the dimensionality of the data [50] (red layers in Figure 1). We used max-pooling as it has been proved the best approach for natural language processing tasks [49]. Finally, these two vectors produced by the convolutional layers are concatenated into a vector of $B \times 400$ that contains the information extracted from both convolutions.

After the concatenation, the vector is fed two times to a pair composed by a fully connected layer with ReLu activation followed by a dropout [51] rate of 0.5374 to prevent the overfitting of the network (green and grey layers in Figure 1). The fully connected layers have 849 and 282 neurons, respectively. Finally, we applied a logarithmic softmax that computes the probability distribution of each label for the tweet (purple in Figure 1). We used the categorical cross-entropy loss function to train this model as it supports multi-class classifications, and it is the most common function for this type of problems. For the optimiser, we used Adam, with the default parameters, as it is the state-of-the-art in optimisations for Deep Learning models as it slightly outperforms the others. The hyperparameters values and the size of the convolutions and the fully connected layers were decided after running a hyperparameter optimisation algorithm.

Finally, we proposed the use of traditional machine learning algorithms as the baseline for the experiments. We decided to use algorithms that are not language focused on proving that language-focused models can classify users and tweets better by focusing on language characteristics and that the tweet elements are not the main features to classify them as this baseline does not obtain good performance even when they are in the tweets. For the experiments conducted in this article, we used four different algorithms: K-Nearest Neighbour [52], Multi-Layer Perceptron [53], Random Forest [54] and Support Vector Classifier [55]. Although we conducted the experiments proposed below with all of them, we only reported the results of the one obtaining the best performance, Random Forest in this case. To feed the text data into the classifier, we tokenised the tweets using the BERT-tokenizer explained before and ran the algorithm using the default parameters provided by SciKit.

B. EVALUATION

To evaluate the hypothesis that the organisations have specific languages, we divided the experimentation into two tasks: tweet-level (RQ1) and user-level (RQ2) classification. Although both tasks aim to classify tweets using the model presented above, each of them solves the proposed hypothesis partly. As previously said, this division responds to the need for both to be correctly resolved to claim that organisations have a specific language. Thus, on the one hand, the proper classification of the tweet-level task allows us to identify that the languages among the five selected organisations are distinct. However, on the other hand, good results in the user-level task show us that the specific language is shared among all the organisation members.

The tweet-level experiment aims to analyse if the languages from the tweets of the different companies are distinguishable by the employed models. Therefore, we balanced the total tweets from each organisation for this experiment, considering the one with the least amount of tweets (390,985 tweets per organisation). Therefore, classifying the tweets with better results than the baseline will mean that there is a common language among the members of the same organisation.

Furthermore, the user-level experiment aims to clarify whether the language detected by the employed models is independent of users and corresponds to the language of a specific organisation. For this experiment, we grouped the tweets produced by each user to avoid having tweets of the same user in the training and testing subsets. To balance the tweets per user and the users per organisations, we took 100 tweets from each user and selected the minimum number of users specified by the organisation with fewer users.

In order to evaluate the used models in both of the tasks, we split the datasets into three different subsets. The training subset represents 70% of the whole data, the validation subset represents 10%, and the testing subset represents the remaining 20%. For the tweet-level experiment, the splits are done randomly, taking into account all available tweets. Instead, the split in the user-level experiment is done by each user, so all the tweets produced by one user are in one subset.

We divided our datasets into three subsets because we employed early stopping to stop the training when it starts to over-fit. After each training iteration, the model is tested with the validation set, obtaining a loss metric and saving the models actual state. Each iteration that loss is compared with the previous one, and when it increases instead of decreasing a limited amount of times (2 in our experiments), the training phase is stopped. Then the test subset is used to test the model's performance as it has remained unseen until this phase.

We also conducted an ablation study on the tweets. This study helped us analyse the importance of each element in the classifications tasks and demonstrate that the organisation's specific language is independent of these elements. Tweets are composed of many elements that help users interact or add more information to the tweet. The elements analysed in this study will be:

- Mention: mentions are used to name other users from Twitter using their usernames (@username). In our analysis, members of the same organisation will usually mention similar users, adding them to the specific language of the organisation.
- Hashtag: hashtags are used to index keywords or topics on Twitter using the key character # (#hashtag). Members of the same organisations will probably use the same hashtags to talk about company topics introducing them to the specific language of the organisation.

- URL: URLs are used to link to external pages as usual. Organisations could use this element to link to similar pages helping to create the specific language of the organisation. The problem regarding URLs in Twitter is that the OSN shortens them, so the name is changed.

We created a dataset for each combination of none, one or two elements for evaluating each element. Then, we followed the same methodology explained in this section to calculate the model's performance for each dataset. The results obtained for the ablation study can be seen in tables 3a and 3b and will be discussed in the next section.

Furthermore, we realised that several tweets in our dataset might contain the name of the organisation and, therefore, the labels in which the model must classify the tweets. To ensure that the model does not learn the names of the organisations to classify the tweets, we replaced them with a unique token. The results of these experiments are shown in Tables 3a and 3b.

We computed the macro F1-score, precision, recall, and accuracy for the experiments conducted in this manuscript. These metrics allow us to know how accurate our classifier is and how robust it is to errors. In the Tables 2, 3a and 3b we reported the metrics for the experiments conducted in the main experiments and the ablations studies. All the experiments ran in an NVIDIA QUADRO RTX 8000.

TABLE 2. Results of the tweet-level and user-level experiments with unprocessed tweets.

	Experiment			
	Tweet-level		User-level	
Model	F1-Score	Accuracy	F1-Score	Accuracy
Random Forest	33,96	34,35%	53,06	56,87%
Bert	**75,66**	76,17%	**87,95**	**87,95%**
Multi-CNN	69,83	69,72%	70,74	71,29%

V. DISCUSSION

After running the general experiments and analysing the results, we can ensure that we found the answer to the key research questions proposed in the first section of the manuscript RQ1 and RQ2. As we can see in Table 2, the BERT model achieved more than 75 points in macro F1-score in both experiments and achieved substantially better results than the Random Forest Classifier, which does not use the language features. Based on the accuracy and the F1 score obtained in the experiments, we can assure that most users use a language with more similarities with their peers than with the rest. We also wanted to analyse if the classifier relies on mentions, hashtags or URLs to classify the tweets.

To analyse how the classification between the different organisations has been, we have extracted a confusion matrix from one experiment carried out at tweet level (Figure 2). We can see that E is the organisation in which most errors are centralised when predicting the others. We hypothesise that this may be because E is the largest organisation of all, and it may be that its users usually variate in their language or talk about very diverse topics that attract the classification of users

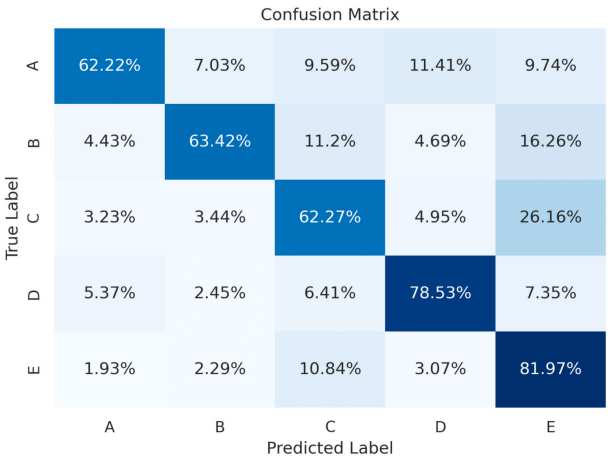

FIGURE 2. Confusion matrix of tweet-level experiments with the Multi-CNN model.

from other organisations. On the other hand, organisation A is the one that accumulates more errors, so it could mean that their language is not as specific as we thought or that in the tweets selected for the test part, the specific language was not as used as in the train tweets, nevertheless, the results are promising.

The experimentation was divided into two tasks to answer the key research questions proposed and ensure that the language used was common to all the organisation members. First, the tweet-level detection task allows us to analyse the tweets individually and detect a language common to all, thus answering RQ1. However, with this approach, the model may focus on specific user styles and forget about the language's style as a whole. To alleviate this problem, the task of user classification was posed. In this second task, the tweets are grouped by user, so the model analyses them together, looking for similarities between users and looking for a common language between them, therefore answering RQ2.

Analysing the ablation study conducted to the tweets, which results are presented in Tables 3a and 3b, we can see how the different combinations of tweet elements affect the classification. Both tweet-level and user-level experiments follow the same importance scale for the tweets' elements with subtle differences. Comparing the model's performance when using only one element in the tweets, we can see that mention is the most important for the classification. This element is used to cite accounts or follow conversations with other users, thus creating higher homophily bias. In our opinion, mentions contribute significantly to the classification because users of the same organisation talk to each other or quote the same accounts when tweeting. URLs and hashtags are also helping the model achieve better results, but as we can see in the presented tables, they need to be combined with others to improve the performance. Hashtags are used to attach tweets to conversations or topics, and the words used in the hashtags may help the classifier detect words in the specific language of an organisation. URLs are used to link

TABLE 3. Results of the experiment conducted in the ablation study performed on the elements of the tweets in (a) tweet-level experiment and (b) user-level experiment.

Elements used	Bert		Multi-CNN		Random Forest	
	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy
None - No orgs names	49,48	50,38%	**57,38**	**57,45%**	33,34	33,83%
None with orgs names	**62,37**	**63,44%**	57,75	57,91%	33,61	34,03%
Hashtag	60,61	61,39%	**62,27**	**62,30%**	34,22	34,55%
Mention	**72,35**	**73,20%**	67,01	66,97%	34,37	34,73%
Url	**64,67**	**65,70%**	57,24	57,38%	33,26	33,64%
Hashtag-Mention	**74,65**	**75,37%**	70,87	70,82%	34,71	35,06%
Hashtag-Url	**65,07**	**65,79%**	60,87	60,92%	33,40	33,78%
Mention-Url	**72,80**	**73,67%**	66,22	66,08%	33,61	34,03%
All	**75,66**	**76,17%**	69,83	69,72%	33,96	34,35%

(a) Tweet-level experiment.

Elements used	Bert		Multi-CNN		Random Forest	
	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy
None - No orgs names	**80,93**	**80,93%**	63,51	65,88%	45,63	51,37%
None with orgs names	**84,66**	**84,66%**	67,27	68,71%	49,87	54,95%
Hashtag	**83,29**	**83,29%**	66,24	67,76%	49,48	54,40%
Mention	**86,85**	**86,85%**	69,71	70,59%	48,88	54,50%
Url	**84,11**	**84,11%**	63,52	65,41%	45,63	51,37%
Hashtag-Mention	**87,12**	**87,12%**	62,18	62,59%	50,65	55,22%
Hashtag-Url	**86,58**	**86,58%**	64,26	64,71%	48,87	54,12%
Mention-Url	**85,75**	**85,75%**	66,42	67,53%	47,47	52,47%
All	**87,95**	**87,95%**	70,74	71,29%	53,06	56,87%

(b) User-level experiment.

external webs with the possibility that several members of the same organisations link the same web pages.

Regarding the experiments conducted with two elements on the tweets, it can be seen that the hashtags with the mentions perform better than the other combinations in both experiments. This is because this combination involves two elements that may represent keywords for the model. For example, if a user is always mentioned by members of the same organisation or a hashtag is continuously used by these members, those words would represent a part of the organisation's specific language.

The ablation study supports the idea that the language of the organisations is not only composed of the tweet elements. Although, using the elements for the classification enhances the performance of the model. The results obtained when replacing all of them, including the organisations' names, demonstrate that the specific language of the organisations is inherent to the elements from Twitter and can be distinguished without them.

Going back to the proposed hypothesis that organisations may have specific languages that their members use on Twitter, we could say that our research strongly supports it by classifying both tweets and users under different conditions with good results. This finding may have a theoretical impact on future researches related to Twitter attack prevention by generating an attention focus on the language of communities and organisations on Twitter in addition to the language of individuals. Furthermore, through the dataset published in the manuscript, we aim to contribute to the implementation or refinement of systems that aim to detect malicious users before they attack organisations on social networks.

VI. CONCLUSION AND FUTURE WORK

In this work, we proposed that language may represent an attack vector usable by malicious users to conduct impersonation or infiltration attacks against organisations on Twitter. For its verification, we gathered tweets from members of five different organisation with a presence on Twitter. Subsequently, we divided the experimentation into two tasks to ensure that the classification was being made by language and not by other factors, such as the style of a unique person. For the experiments, we employed state-of-the-art models for text

classification that obtained almost 75% of accuracy in both tasks.

We also conducted an ablation study to analyse which of the elements present on a tweet offer more information about the organisation to whom its author is a member. Besides confirming that mentions and hashtags are the most crucial elements to classify tweets, we confirmed that the selected organisations' specific languages might be classified without relying on these elements. Furthermore, the successful results obtained by our algorithm when we changed all the tweet elements and the name of the organisations for tokens show that members of the same organisation share an inherent writing style that is independent of hashtags, mentions and URLs. Thus, with the experiments conducted in this article, we can ensure that organisations have specific languages on Twitter.

Therefore, we aim to generate artificial tweets based on actual tweets from a single organisation to test if the classifier can differentiate them from real tweets in our future work. We also want to verify if the algorithm proposed in this manuscript can classify the artificial tweets in the correct organisation. The final objective is to train a model to detect tweets artificially created to infiltrate organisations and stop possible attacks from malicious users.

ACKNOWLEDGMENT

The authors would like to gratefully acknowledge the support of NVIDIA Corporation for the donation of the hardware used in this research and the Weight and Biases Team for their service and support.

REFERENCES

[1] S. Rathore, P. K. Sharma, V. Loia, Y.-S. Jeong, and J. H. Park, "Social network security: Issues, challenges, threats, and solutions," *Inf. Sci.*, vol. 421, pp. 43–69, Dec. 2017.

[2] D. Pacheco, P.-M. Hui, C. Torres-Lugo, B. T. Truong, A. Flammini, and F. Menczer, "Uncovering coordinated networks on social media," 2020, *arXiv:2001.05658*. [Online]. Available: https://arxiv.org/abs/2001.05658

[3] D. Pacheco, A. Flammini, and F. Menczer, "Unveiling coordinated groups behind white helmets disinformation," in *Proc. Companion Proc. Web Conf.*, Apr. 2020, pp. 611–616.

[4] S. Zannettou, T. Caulfield, E. De Cristofaro, M. Sirivianos, G. Stringhini, and J. Blackburn, "Disinformation warfare: Understanding state-sponsored trolls on Twitter and their influence on the web," in *Proc. Companion World Wide Web Conf.*, May 2019, pp. 218–226.

[5] *Insights Into Attempts to Manipulate Twitter by State-Backed Entities*, Twitter, San Francisco, CA, USA, 2020.

[6] A. Bessi and E. Ferrara, "Social bots distort the 2016 US Presidential election online discussion," *1st Monday*, vol. 21, no. 7, pp. 1–14, Nov. 2016.

[7] 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. 374–378, Dec. 2019.

[8] G. Enli, "Twitter as arena for the authentic outsider: Exploring the social media campaigns of Trump and Clinton in the 2016 US presidential election," *Eur. J. Commun.*, vol. 32, no. 1, pp. 50–61, Feb. 2017.

[9] M. Grčar, D. Cherepnalkoski, I. Mozetič, and P. K. Novak, "Stance and influence of Twitter users regarding the brexit referendum," *Comput. Social Netw.*, vol. 4, no. 1, pp. 1–25, Dec. 2017.

[10] M. Mora-Cantallops, S. Sánchez-Alonso, and A. Visvizi, "The influence of external political events on social networks: The case of the Brexit Twitter Network," *J. Ambient Intell. Hum. Comput.*, vol. 12, no. 4, pp. 4363–4375, Apr. 2021.

[11] M. Stella, E. Ferrara, and M. De Domenico, "Bots increase exposure to negative and inflammatory content in online social systems," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 49, pp. 12435–12440, Dec. 2018.

[12] A. M. Jamison, D. A. Broniatowski, and S. C. Quinn, "Malicious actors on Twitter: A guide for public health researchers," *Amer. J. Public Health*, vol. 109, no. 5, pp. 688–692, May 2019.

[13] R. Venkatesh, J. K. Rout, and S. Jena, "Malicious account detection based on short URLs in Twitter," in *Proc. Int. Conf. Signal, Netw., Comput., Syst.* New York, NY, USA: Springer, 2017, pp. 243–251.

[14] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annu. Rev. Sociol.*, vol. 27, no. 1, pp. 415–444, 2001.

[15] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.

[16] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8110–8119.

[17] C. Freitas, F. Benevenuto, A. Veloso, and S. Ghosh, "An empirical study of socialbot infiltration strategies in the Twitter social network," *Social Netw. Anal. Mining*, vol. 6, no. 1, p. 23, Dec. 2016.

[18] Y. Jun, R. Meng, and G. V. Johar, "Perceived social presence reduces fact-checking," *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 23, pp. 5976–5981, Jun. 2017.

[19] K. Zarei, R. Farahbakhsh, and N. Crespi, "Deep dive on politician impersonating accounts in social media," in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, Jun. 2019, pp. 1–6.

[20] T. B. N. Hoang and J. Mothe, "Location extraction from tweets," *Inf. Process. Manage.*, vol. 54, no. 2, pp. 129–144, Mar. 2018.

[21] B. Mei, Y. Xiao, H. Li, X. Cheng, and Y. Sun, "Inference attacks based on neural networks in social networks," in *Proc. 5th ACM/IEEE Workshop Hot Topics Web Syst. Technol.*, Oct. 2017, pp. 1–6.

[22] F. Zarrinkalam, M. Kahani, and E. Bagheri, "Mining user interests over active topics on social networks," *Inf. Process. Manage.*, vol. 54, no. 2, pp. 339–357, Mar. 2018.

[23] O. Varol, E. Ferrara, F. Menczer, and A. Flammini, "Early detection of promoted campaigns on social media," *EPJ Data Sci.*, vol. 6, no. 1, pp. 1–9, Dec. 2017.

[24] X. Zhang and A. A. Ghorbani, "An overview of online fake news: Characterization, detection, and discussion," *Inf. Process. Manage.*, vol. 57, no. 2, Mar. 2020, Art. no. 102025.

[25] 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.

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

[27] K. 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, Jan. 2019.

[28] M. Orabi, D. Mouheb, Z. Al Aghbari, and I. Kamel, "Detection of bots in social media: A systematic review," *Inf. Process. Manage.*, vol. 57, no. 4, Jul. 2020, Art. no. 102250.

[29] A. Das, S. Gollapudi, E. Kıcıman, and O. Varol, "Information dissemination in heterogeneous-intent networks," in *Proc. 8th ACM Conf. Web Sci.*, pp. 259–268, 2016.

[30] S. Argamon-Engelson, M. Koppel, and G. Avneri, "Style-based text categorization: What newspaper am I reading," in *Proc. AAAI Workshop Text Categorization*, 1998, pp. 1–4.

[31] N. Potha and E. Stamatatos, "Intrinsic author verification using topic modeling," in *Proc. 10th Hellenic Conf. Artif. Intell.*, Jul. 2018, pp. 1–7.

[32] E. Stamatatos, "Authorship attribution using text distortion," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2017, pp. 1138–1149.

[33] O. Fourkioti, S. Symeonidis, and A. Arampatzis, "Language models and fusion for authorship attribution," *Inf. Process. Manage.*, vol. 56, no. 6, 2019, Art. no. 102061.

[34] A. Rocha, W. J. Scheirer, C. W. Forstall, T. Cavalcante, A. Theophilo, B. Shen, A. R. Carvalho, and E. Stamatatos, "Authorship attribution for social media forensics," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 1, pp. 5–33, Jan. 2017.

[35] P. Bellot, C. Trabelsi, J. Mothe, F. Murtagh, J. Y. Nie, L. Soulier, E. SanJuan, L. Cappellato, and N. Ferro, "Experimental ir meets multilinguality, multimodality, and interaction," in *Proc. 9th Int. Conf. CLEF Assoc.* (Lecture Notes in Computer Science), vol. 11018. New York, NY, USA: Springer, 2018, pp. 267–285.
[36] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2014, pp. 2672–2680.
[37] D. Jin, Z. Jin, Z. Hu, O. Vechtomova, and R. Mihalcea, "Deep learning for text style transfer: A survey," 2020, *arXiv:2011.00416*. [Online]. Available: http://arxiv.org/abs/2011.00416
[38] H. Jhamtani, V. Gangal, E. Hovy, and E. Nyberg, "Shakespearizing modern language using copy-enriched sequence to sequence models," in *Proc. Workshop Stylistic Variation*, 2017, pp. 10–19.
[39] K. Carlson, A. Riddell, and D. Rockmore, "Evaluating prose style transfer with the bible," *Roy. Soc. Open Sci.*, vol. 5, no. 10, Oct. 2018, Art. no. 171920.
[40] M. Fire and R. Puzis, "Organization mining using online social networks," *Netw. Spatial Econ.*, vol. 16, no. 2, pp. 545–578, Jun. 2016.
[41] A. Singh, E. Blanco, and W. Jin, "Incorporating emoji descriptions improves tweet classification," in *Proc. Conf. North*, vol. 1, 2019, pp. 2096–2101.
[42] Z. Chen, S. Shen, Z. Hu, X. Lu, Q. Mei, and X. Liu, "Emoji-powered representation learning for cross-lingual sentiment classification," in *Proc. World Wide Web Conf.*, 2019, pp. 251–262.
[43] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Jun. 2019, pp. 4171–4186.
[44] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune bert for text classification," in *Proc. China Nat. Conf. Chin. Comput. Linguistics*. New York, NY, USA: Springer, 2019, pp. 194–206.
[45] J. Y. Lee and F. Dernoncourt, "Sequential short-text classification with recurrent and convolutional neural networks," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2016, pp. 515–520.
[46] S. Poria, E. Cambria, and A. Gelbukh, "Aspect extraction for opinion mining with a deep convolutional neural network," *Knowl.-Based Syst.*, vol. 108, pp. 42–49, Sep. 2016.
[47] A. Severyn and A. Moschitti, "Twitter sentiment analysis with deep convolutional neural networks," in *Proc. 38th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2015, pp. 959–962.
[48] A. Bilbao-Jayo and A. Almeida, "Political discourse classification in social networks using context sensitive convolutional neural networks," in *Proc. 6th Int. Workshop Natural Lang. Process. Social Media*, 2018, pp. 76–85.
[49] Y. Zhang and B. C. Wallace, "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification," in *Proc. 8th Int. Joint Conf. Natural Lang. Process.*, vol. 1, 2017, pp. 253–263.
[50] Y.-L. Boureau, J. Ponce, and Y. LeCun, "A theoretical analysis of feature pooling in visual recognition," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 111–118.
[51] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[52] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *Amer. Statist.*, vol. 46, no. 3, pp. 175–185, 1992.
[53] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," Jolla Inst. Cogn. Sci., California Univ San Diego La, San Diego, CA, USA, Tech. Rep. 8506, 1985.
[54] S. Shalev-Shwartz and S. Ben-David, *Decision Trees*. Cambridge, U.K.: Cambridge Univ. Press, 2014, pp. 212–218.
[55] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

RUBÉN SÁNCHEZ-CORCUERA received the bachelor's degree in computer science and the master's degree in computer engineering from the University of Deusto, in 2017 and 2018, respectively. He is currently pursuing the Ph.D. degree with the DEUSTEK Research Group, Faculty of Engineering, University of Deusto. His main research interests include analysis of online social networks using natural language processing and graph-based methods.

ARKAITZ ZUBIAGA is currently a Lecturer with the Queen Mary University of London, UK, where he leads the Social Data Science Laboratory. He has published over 100 articles (more than 30 journal articles) in interdisciplinary areas, including social data science, computational social science, and natural language processing. He is also a regular SPC member of top conferences, including WWW, ICWSM, CHI, CSCW, ACL, EMNLP, IJCAI, and AAAI. He serves as an Academic Editor for *Online Social Networks and Media* and *PeerJ Computer Science and Information*, and an Editorial Board Member for *Information Processing and Management*.

AITOR ALMEIDA received the Ph.D. degree in computer science from the University of Deusto. He is currently a Researcher and the Project Manager with the DeustoTech Institute, Faculty of Engineering, University of Deusto. His research interests include the analysis of the behavior of the users in intelligent environments, the application of artificial intelligence for smart health and the study of the users' activity, and discourse on social networks.