JOURNAL OF SOCIAL COMPUTING
ISSN 2688-5255 04/06 pp46−61
Volume 4, Number 1, March 2023
DOI: 10.23919/JSC.2023.0001

Manipify: An Automated Framework for Detecting Manipulators in Twitter Trends

Soufia Kausar, Bilal Tahir*, and Muhammad Amir Mehmood

Abstract: The rapid adoption of online social media platforms has transformed the way of communication and interaction. On these platforms, discussions in the form of trending topics provide a glimpse of events happening around the world in real-time. Also, these trends are used for political campaigns, public awareness, and brand promotions. Consequently, these trends are sensitive to manipulation by malicious users who aim to mislead the mass audience. In this article, we identify and study the characteristics of users involved in the manipulation of Twitter trends in Pakistan. We propose "Manipify"—a framework for automatic detection and analysis of malicious users in Twitter trends. Our framework consists of three distinct modules: (1) user classifier, (2) hashtag classifier, and (3) trend analyzer. The user classifier module introduces a novel approach to automatically detect manipulators using tweet content and user behaviour features. Also, the module classifies human and bot users. Next, the hashtag classifier categorizes trending hashtags into six categories assisting in examining manipulators behaviour across different categories. Finally, the trend analyzer module examines users, hashtags, and tweets for hashtag reach, linguistic features, and user behaviour. Our user classifier module achieves 0.92 and 0.98 accuracy in classifying manipulators and bots, respectively. We further test Manipify on the dataset comprising 652 trending hashtags with 5.4 million tweets and 1.9 million users. The analysis of trends reveals that the trending panel is mostly dominated by political hashtags. In addition, our results show a higher contribution of human accounts in trend manipulation as compared to bots.

Key words: trend manipulation; bot classification; user analysis; text classification; manipulator detection

1 Introduction

Online social media platforms have emerged as a key source of information and socializing during last decade. These platforms strive to maximize user engagement through rapid information dissemination to information savvy users. In this regard, Twitter—a micro-blogging platform—provides real-time trends of the most discussed topics in the trending panel[1]. Due to the extensive reach, such trends have enabled journalists and business analysts to explore breaking news, predict candidate popularity, and to evaluate product reviews[2]. In addition, a survey reports that 74% of Twitter users utilise this platform as a source of daily news while 34% of these users focus on trending topics for this purpose[3]. On one hand, Twitter trends are being used to detect breakthrough events, product marketing, and crisis management[4]. On the other hand, these trends are subjected to manipulation by malicious users to spread false narratives[5].

Recent research reveals that Twitter trends can easily be manipulated by using a small number of automated accounts[6]. A new business has emerged where companies are selling Twitter trends "manipulation as a service" to produce false trends[7]. A survey reveals that 23%–27% conversations on Twitter related to politics during US elections 2016 are carried out by bot accounts[8]. Also, another research identifies 40% bot users disseminating information related to COVID-19

• Soufia Kausar, Bilal Tahir, and Muhammad Amir Mehmood are with the Al-Khawarizmi Institute of Computer Science, University of Engineering and Technology, Lahore 54890, Pakistan. E-mail: soufiamirza5@gmail.com; {bilal.tahir, amir.mehmood}@kics.edu.pk.
* To whom correspondence should be addressed.
Manuscript received: 2022-11-16; revised: 2023-02-23; accepted: 2023-04-13

on Twitter[9].

In general, researchers have focused on examining trend manipulation by analyzing the activity of human and bot accounts[10]. Also, the pattern of deletion of tweets related to a trend is studied to detect the possibility of manipulation[11]. Moreover, limited number of trending hashtags are manually examined for manipulation[5]. However, these approaches have three major limitations. First, only bot accounts cannot be labelled as manipulators as human accounts are also involved in trend manipulation[5]. Second, manipulators do not necessarily delete tweets after creating a trend. Finally, the manual analysis of constantly emerging new trends is not possible. Moreover, proposed techniques for the identification of spam, bot, fake, compromised, and cloned accounts are not extendable for manipulators due to the dissimilarity between their behaviours.

In this paper, we propose a novel framework of "Manipify" to automatically identify and examine the manipulators in Twitter trends. Our framework consists of three major modules: (1) user classifier, (2) hashtag classifier, and (3) trend analyzer. The first module of the user classifier identifies manipulators leveraging our developed Manipulator Detection Dataset (MT-Dat) and five proposed features related to content and user behaviour. In addition, the user classifier also identifies bot and human accounts using user profiles, behaviour, and activity features. For the classification of bot and manipulator accounts, we compare the performance of six machine learning classifiers to determine the best-performing model. We report the highest accuracy of 0.92 for manipulator detection with Logistic Regression (LR) and 0.98 for bot identification with the Decision Tree (DT) classifier. In addition, the hashtag classifier categorizes hashtags into six classes: (1) politics, (2) sports, (3) religion, (4) campaign, (5) entertainment, and (6) military using our Hashtag Classification Dataset (Ha-Dat) of 2384 hashtags. Finally, the trend analyzer examines the trends for language distribution of tweets, topics and reach of trending hashtags, and the behaviour of manipulators and bots. We conduct a case study on popular hashtags originating from Pakistan, given their susceptibility to being exploited for the 5th generation war, political purposes, and economic benefits. For this purpose, we build a PK-Trends dataset containing 652 trending hashtags from Pakistan, 5.4 million tweets, and 1.9 million users. Specifically, we collect trending hashtags and their related tweets for one week in November 2020, December 2020, and January 2021 with a gap of five weeks to minimize the sampling bias. Our major contributions and key findings are summarized as follows.

- We introduce a novel Manipify framework to automatically detect users involved in the manipulation of Twitter trends with the accuracy of 0.92. In addition, our framework identifies bot accounts and categorizes trends into six categories for comprehensive analysis of trending hashtags.
- Our analysis of trending hashtags from Pakistan reveals that political and campaign hashtags dominate the trending panel with 32%–42% and 16%–32% hashtags, respectively. On the contrary, only 1%–14% hashtags are from the categories such as sports, entertainment, and religion.
- Our analysis of users highlights that our dataset of local trends from Pakistan contains 2.58×10^5 (39%) and 5.6×10^4 (8.3%) bot and manipulator accounts, respectively. Moreover, 56.7% of manipulator accounts are humans.
- We note that 20% manipulator accounts manipulate more than one hashtag. Zooming into such users, human manipulators associate themselves with hashtags related to only one category. On the other hand, a bot manipulator manipulates hashtags of different categories.

The rest of the paper is structured as follows. Section 2 presents the related work and Section 3 introduces the developed datasets. In Section 4, we describe the Manipify framework while its evaluation is presented in Section 5. Section 6 presents the analysis conducted for Twitter trends in Pakistan using Manipify. Finally, we conclude our work in Section 7.

2 Related Work

Recently, social media analysis had been adopted to perform topic-based sentiment analysis, public opinion-mining, emotion analysis, crime monitoring, and spam detection[12]. In addition, social media users were examined for malicious user identification, location inference, and spammer identification[13].

In the last decade, researchers had focused on examining Twitter trends due to their impact on society. For instance, a real-time system was developed for the classification of trending topics into news, current events,

memes, and commemoratives[14]. Zubiaga et al.[14] used features from the tweets text and their metadata for the classification of trends. In addition, Zhang et al.[15] investigated the possibility of trend manipulation on Twitter. They experimented with features such as popularity and coverage for Twitter trending topics to inspect features that contribute more towards the prediction of trends. Their analysis also indicated the presence of malicious and spam users manipulating Twitter trends. Similarly, Khan et al.[16] proposed a real-time trend detection method by analyzing a stream of tweets. They used statistical information retrieval methods to extract important terms. Furthermore, the first large-scale study on manipulated/fake Twitter trends was conducted by Elmas et al.[11] They uncovered the fact that nearly 20% of the global Twitter trends were a result of manipulation.

Twitter users are key to disseminating information and creating trends. Studying the characteristics of these users, Motamedi et al.[17] conducted a detailed study on two snapshots of elite user accounts present on Twitter. They investigated features of elite users such as a change in followers, followees, and rank over time. Also, the findings showed that graph relation between elite users formed 14–20 communities. Similarly, an analysis of one million Twitter users was carried out to analyze the behaviour of demographic groups[18]. The analysis of demographic attributes including gender, ethnicity, and account age of one million Twitter users highlighted that various demographic groups show differences in behaviour. In addition, Yaqub et al.[19] conducted sentiment analysis of two political candidates during 2016 US presidential elections. The analysis showed that Trump received more positive sentiment as compared to Hillary Clinton. In addition, they analyzed tweets of one million Twitter users to identify their opinion. Their findings revealed that existing opinions were re-shared using the retweet feature instead of building new opinions and arguments. Furthermore, political micro-influencers were identified and examined in the Twitter space of Pakistan[20]. The longitudinal analysis of 526 micro-influencers from 2018 to 2020 revealed that 40% accounts no longer exist after two years. Also, 22% of micro-influencers in 2018 became macro-influencers in 2020. Finally, we previously proposed a framework of "Push-To-Trend" to automatically identify the trend promoters[13]. Push-To-Trend used features of the number of total tweets, duplicate tweets, overlapping n-grams, and peak-to-mean ratio for identification of trend promoters. The framework was also utilized to detect 1.47×10^5 trend promoters related to 602 hashtags in a large-scale Urdu tweets repository of Anbar[21].

The automatic understanding of hashtags is a challenging task because hashtags are inconsistent and lack standard vocabulary[22]. Previously, the researchers had adopted the labour-intensive and time-consuming path for labelling the trending hashtags. In this regard, Romero et al.[23] manually categorised hashtags into eight pre-defined categories to examine the patterns of information dissemination. Jeon et al.[24] experimented to build a hashtag recommendation system after their topic classification using Term Frequency Inverse Document Frequency (TF-IDF) lexical features. Another algorithm was proposed for hashtag classification after combining lexical and pragmatic features[25]. The pragmatic features were related to user profiles such as the number of followers or followees. In addition, open-source content like Wikipedia and Open Directory was utilized for the classification of hashtags[26]. For example, Ferragina et al.[26] used the Wikipedia graph to devise the Hashtag-Entities (HE) graph which represented the semantic relation between hashtags and their entities. However, this approach was limited to the hashtags that were available in the Wikipedia graph only. Moreover, HashCat framework was proposed for the classification of multilingual hashtags[27].

The literature review revealed that the research community had focused on exploring the possibility of manipulation of Twitter trends but no technique had been presented for the automatic detection of manipulators. To the best of our knowledge, only one framework of Push-To-Trend was proposed to identify users promoting the trending hashtag. Push-To-Trend focused on detecting the users promoting the hashtags. In contrast, we propose a machine learning based model for automatic detection of manipulators generating fake Twitter trends. Moreover, we build an analyzer module for a comprehensive analysis of Twitter trends.

3 Dataset

In this section, first, we focus on building datasets to detect manipulators and bots. Next, we describe the

process of developing a dataset for hashtag classification. Finally, we discuss our PK-Trends dataset to study Pakistan Twitter trends.

3.1 Manipulator detection dataset—MT-Dat

The development of an automated system for the identification of manipulators requires a gold-standard labelled dataset. The absence of such a dataset motivates us to build Manipulator Detection Dataset (MT-Dat) with "manipulator" and "organic" users. For this purpose, first, we collect 5.4 million tweets posted by 1.9 million users related to 652 trending hashtags from Pakistan. In addition, we deliberately collect the tweets related to trending hashtags for three weeks in November 2020, December 2020, and January 2021 with a gap of five weeks each to minimize sampling bias. Details of the complete dataset will be described shortly.

For labelling, we randomly select users and human annotators examine the velocity, volume, and content similarity of tweets posted by users related to the trending hashtags for labelling[28]. It must be noted that a user posting tweets related to multiple hashtags is considered as separate samples for labelling. In particular, human annotators are furnished with tweets related to trending hashtags, the time stamp of tweets, and the trending time of hashtags along with guidelines for labelling. The guidelines provided to assign the label of organic and manipulator users are as follows.

- **G1:** We are interested in users responsible for creating the trend. Therefore, only analyze the tweets and user behaviour before the trend is first seen in the trending panel.
- **G2:** Users posting a large number of tweets related to trending hashtags in a small amount of time to increase the velocity are labelled as manipulators.
- **G3:** Assign the label of the manipulator to the user posting multiple tweets with similar content to increase the volume.
- **G4:** Users posting tweets containing only trending hashtags without any additional information are manipulators.

Two annotators manually label the 1010 randomly selected samples according to these guidelines. In MT-Dat, 510 users are labelled as manipulators and 500 are labelled as non-manipulators. Also, these samples are related to 225 trending hashtags. Table 1 shows the statistics of our dataset. We report the mutual agreement of 98% to the labels assigned by two annotators. In addition, we inspect 17 871 tweets related to 1010 samples where 7816 (44%), 6090 (34%), 2457 (14%), and 1506 (8%) tweets are posted in Urdu, English, unknown, and other languages, respectively. This diverse representation of languages highlights that our dataset captures the rich and varied features of users posting content in different languages for efficient analysis of real-world content.

Table 1 Statistics of MT-Dat.

Category	Number of users	Percentage of users (%)
Manipulator	510	50.4
Non-manipulator	500	49.6
Total	1010	100

3.2 Bot detection dataset—BT-Dat

Next, we use three publicly available datasets to develop our Bot Detection Dataset (BT-Dat). The first labelled dataset of midterm-2018 dataset contains information of users and tweets from the US midterm elections 2018[29, 30]. Annotators label the user as human if they are actively involved in any political discussions. To label bot accounts, features of tweet time and account creation time are manually observed. The midterm-2018 dataset contains a total of 50 538 user accounts from which 42 446 are bot and 8092 are human accounts. We only include 11 908 bots from midterm-2018 in our dataset. Due to the insufficient number of human accounts in midterm-2018, we also use the dataset of celebrity-2019[31]. This dataset contains 4589 human accounts belonging to prominent public figures. In addition, the third dataset of Cresci-2017 contains 7543 bot and 3474 human accounts. The accounts in the dataset are labelled by the Crowdflower contributors[¤]. For human account labelling, annotators contacted random Twitter users and asked a question in the natural language. Accounts that answered questions properly are labelled as human accounts. Moreover, Cresci-2017 contains three classes of bots: (1) traditional bots, (2) fake followers, and (3) social spambots[32]. We combine users in all three datasets to develop a comprehensive Bot Detection Dataset (BT-Dat) that contains 35 606 labelled Twitter accounts. Table 2 shows statistics for the four bot detection datasets.

[¤]http://faircrowd.work/platform/crowdflower

Table 2 Statistics of BT-Dat.

Dataset	Number of accounts	Number of bots	Percentage of bots (%)	Number of humans	Percentage of humans (%)
Midterm-2018	20 000	11 908	60	8092	40
Celebrity-2019	4589	0	0	4589	100
Cresci-2017	11 017	7543	68	3474	32
BT-Dat	35 606	19 451	55	16 155	45

3.3 Hashtag classification dataset—Ha-Dat

As such, the hashtags contain symbols, names of organizations, people, or events joined without using any space[22]. A labelled hashtags dataset is required to understand and classify these hashtags into their respective categories. To this goal, we develop a labelled hashtag dataset by manually annotating trending hashtags into six categories of (1) politics, (2) sports, (3) religion, (4) campaign, (5) entertainment, and (6) military. We collected trending hashtags from Twitter using the Twitter Application Programming Interface (API)[33] from 25 July 2018 to 6 August 2021. Using the definitions of hashtag categories[27], two annotators manually labelled randomly selected 2384 hashtags. We fetch tweets related to these hashtags in English and Urdu language. The hashtags that contain tweets in the English language only are referred to as "English hashtags". Similarly, hashtags that contain Urdu language tweets only are "Urdu hashtags". Finally, "English-Urdu hashtags" are those that contain tweets in both languages. The detailed statistics of our dataset are shown in Table 3.

3.4 PK-Trends dataset

The manipulation of Twitter trends is a global phenomenon which has also adversely affected the digital space of Pakistan[34]. Twitter is the 5th most popular social media platform in Pakistan with 4.65 million users[35]. In recent time, mainstream media have repeatedly reported the promotion of fake trends in the trending panel of Pakistan[36]. In this context, companies and social media groups are also identified in Pakistan providing paid services to create fake Twitter trends[37]. Furthermore, Twitter trends in Pakistan have also been subjected to the 5th generation warfare[38]. Given these factors, we analyze trends in Pakistani Twitter as a case study and perform multi-facet analysis of tweets and users.

First, we build PK-Trends dataset by collecting trending hashtags in Pakistan from the online service of GetDayTrends[†] which provides a list of trending hashtags on an hourly basis. We create three datasets by fetching trending hashtags for one week in November 2020 (PK-Nov-20), December 2020 (PK-Dec-20), and January 2021 (PK-Jan-21). We deliberately take these samples with a gap of five weeks to explore trending hashtag dynamics. Also, this approach assists in minimizing the sampling and seasonal bias. In addition, we define a time window to fetch tweets related to the trending hashtags. This time window includes tweets from one day before to one day after the hashtag is seen in the trending panel. Moreover, the Python language library of Twint[‡] is used to fetch the tweets for a three-day time window. We note that Twint could not scrap "retweets". Therefore, cumulatively, PK-Trends dataset contains 5.4 million "original" tweets posted by 1.9 million users. Furthermore, PK-Nov-20, PK-Dec-20, and PK-Jan-21 contain 281, 184, and 187 unique trending hashtags, respectively. Table 4 shows statistics of PK-Trends dataset.

We analyze trending hashtags and related tweets in PK-Trends for seven days with 2 h bins for all datasets. We observe a periodic pattern in the number of unique trending hashtags with a peak around midday (11: 00 AM–3: 00 PM) consistent with prior studies[39]. This observation highlights that users discuss more unique hashtags during the daytime. On average, PK-Nov-20, PK-Dec-20, and PK-Jan-21 contain tweets related to 104 (36.6%), 64 (34.1%), and 42 (21.7%) trending hashtags in each bin, respectively. Also, the PK-Trends dataset contains tweets posted in multiple languages. To study the natural language distribution, we use the meta-information of tweet language provided by Twitter. We found that English is the most frequent language used in PK-Trends dataset with 3.2 million tweets. Whereas, only 0.48 million tweets are posted in the Urdu language. It has 0.67 million tweets marked as unknown language and 1.1 million tweets from other languages.

4 Manipify—Framework

In this section, we describe the architecture of our

[†]https: //getdaytrends.com/pakistan/
[‡]https: //github.com/twintproject/twint

Table 3 Statistics of Ha-Dat.

Language	Number of hashtags							
	Politics	Sports	Religion	Campaign	Entertainment	Military	Other	Total
English	292	120	77	185	117	102	44	937
Urdu	212	18	102	105	35	38	56	566
English-Urdu	359	89	75	160	75	29	94	881
Total	863	227	254	450	227	169	194	2384

Table 4 Statistics of PK-Trends.

Dataset	Attribute							
	Time period	Unique trends	Unique hashtags	Local hashtags	Global hashtags	Unique keywords	Unique users	Unique tweets
PK-Nov-20	08–15 Nov 2020	1542	281	236	45	1261	1 359 406	2 990 850
PK-Dec-20	13–20 Dec 2020	1454	184	161	23	1270	554 513	2 045 448
PK-Jan-21	21–28 Jan 2021	1391	187	129	58	1204	75 628	458 799

proposed Manipify framework. First, we explain the user classification modules. Next, we discuss the methodology of hashtag classification. Finally, we present the trend analyzer module proposed to understand the various dynamics of users.

4.1 Manipulator classification

Figure 1 shows the architecture diagram of Manipify. The classification of manipulators requires distinct features related to users to train the machine learning model. In the literature, there is no study available which automatically identifies the manipulators using such features. Hence, we design five features related to users which are (1) the number of total tweets by a user (Tweets), (2) the number of tweets before trend time ($\text{Tweets}_{\text{before}}$), (3) average time between consecutive tweets after trend time ($\text{Time}_{\text{after}}$), (4) average time between consecutive tweets before trend time ($\text{Time}_{\text{before}}$), and (5) content similarity score ($\text{Sim}_{\text{score}}$). As such, the volume and velocity of tweets containing a hashtag are key factors to determine the trending hashtags[28]. In literature, the manual analysis of manipulators reveals that they post a large number of tweets (Tweets) using a hashtag to create a trend[5]. Particularly, it is observed that these users post tweets before trend time ($\text{Tweets}_{\text{before}}$) as "organic user" generally uses the hashtag after it is seen in the trending panel. In addition, the velocity of tweets containing a hashtag is the key factor to determine the

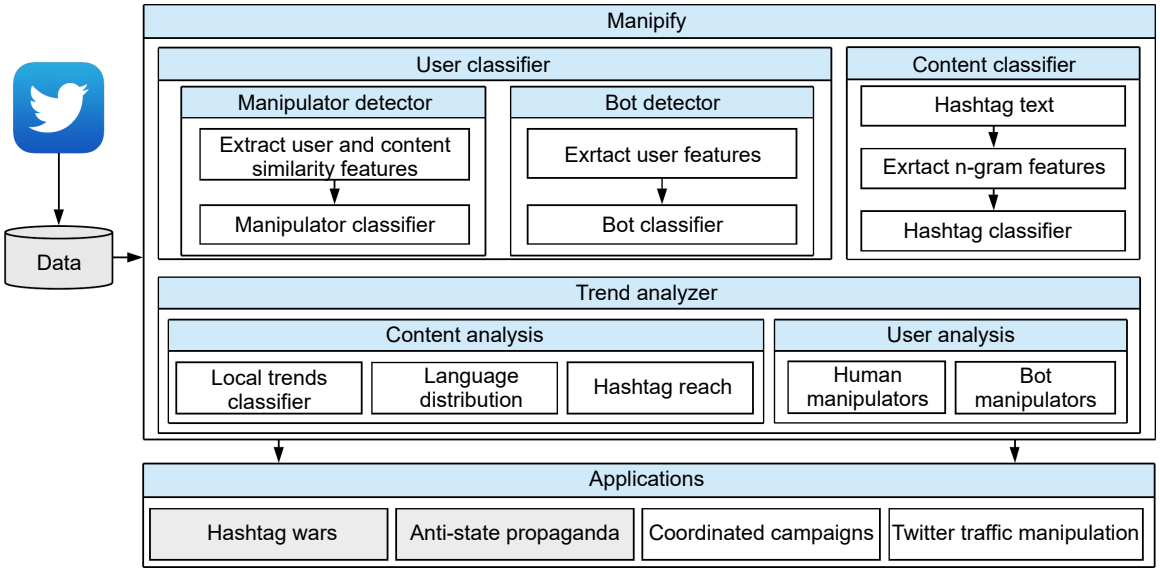

Fig. 1 Architecture diagram of our research methodology.

trend. Therefore, we consider the average time between tweets posted by user before ($\text{Time}_{\text{before}}$) and after ($\text{Time}_{\text{after}}$) trending time of a hashtag. We calculate the velocity of tweets before and after trending time as we believe manipulators limit the activity after trending which impacts the overall value of tweets velocity.

Finally, we calculate the similarity score ($\text{Sim}_{\text{score}}$) of tweets posted by a user with the aim of identifying the manipulators posting similar tweets to increase the volume. In addition, posting and deleting a large number of tweets using the same hashtag is a violation of Twitter's platform policy[40]. To calculate the content similarity score, we make use of natural language processing techniques. The similarity score of all tweets posted by users related to a hashtag is computed by using the overlapping n-grams in tweets. It is worth noting that the comparison of all possible n-grams of tweets with a length greater than one is done for score calculation. Let a user (U) post m tweets related to a hashtag (H). We calculate the similarity score ($\text{Sim}_{\text{score}}$) of users concerning hashtags by computing the ratio of overlapping n-grams and total n-grams of all tweets posted by users using Eq. (1).

$$\text{Sim}_{\text{score}}(U) = \frac{\sum_{w=2}^{m}(\text{n-gram}_w \times \text{freq(n-gram)})}{\text{Total n-grams}} \quad (1)$$

It should be noted that we only utilise the original tweets of a user for calculating the classification features and retweets are not considered.

Availing five features of users, we compare the performance of six state-of-the-art classifiers: (1) Logistic Regression (LR), (2) Decision Tree (DT), (3) Gaussian Naive Bayes (GNB), (4) Random Forest (RF), (5) K Nearest Neighbor (KNN), and (6) Support Vector Machine (SVM) for classifying the manipulators[41]. Moreover, we compare the performance of models using metrics of precision, recall, F1-score, accuracy, and Area Under Curve (AUC). Also, the split ratio of 70:30 is used for manipulator classification.

4.2 Bot detection

We examine the user profile, behaviour, and activity features for the classification of bots. Specifically, we compare the efficacy of features for the identification of bot accounts proposed by Efthimion et al.[42] and Yang et al.[43] Efthimion et al.[42] extracted nine binary features related to a user account: (1) user description, (2) URL existing in the description, (3) friend count > 1000, (4) follower count < 30, (5) user geo-location, (6) list count > 0, (7) statuses count > 0, (8) URL existing in profile, and (9) user verified. In general, the presence of URL in profile and description indicates the human account because bot accounts lack such customized information. Also, in bot accounts profile, geo-location is not typically available. Similarly, human users tend to have a lower number of friends. On the other hand, bot accounts have a large number of friends count. In addition, Yang et al.[43] utilized 19 meta and derived features of users for bot classification. These features include the number of tweets, followers, friends, and favourite count of users. Also, three binary features availability of default profile, verified account, and profile background are extracted. In addition, computed derived features are frequency of tweets per year, follower/following growth rate, length of user name, screen name, description, etc. Similar to manipulator classification, we compare the performance of six machine learning classifiers with five evaluation metrics for bot classification with a split ratio of 70:30.

4.3 Content classification

To build a hashtag classification module, we leverage the lexical features of tweets for the classification of hashtags as done by Jeon et al.[24] First, during the pre-processing of tweet text, we remove all symbols, hashtags, and URLs. Next, we extract 1−3 n-grams utilising the statistical technique of Term Frequency Inverse Document Frequency (TF-IDF) from all pre-processed tweets related to a hashtag. Using the extracted TF-IDF features, we train a logistic regression classifier with default parameters for classification. It is crucial to mention here that we train separate TF-IDF vectorizers and models for English and Urdu hashtags. Moreover, we train a series of binary classifiers for each hashtag category using the "one-vs-all" approach due to its better performance compared to multi-class classifiers[44, 45]. We assign the category of the classifier with the highest probability to the hashtag. The hashtag is assigned the label of "Other" if binary classifiers of all categories give the probability less than 0.5. Moreover, we need to calculate the minimum number of required tweets to

accurately classify the topic of a hashtag. In this regard, we use 100 as the minimum number of tweets required to classify the hashtag in accordance with Refs. [24, 46].

4.4 Trend analyzer

Next, we turn our attention to the analyzer module which is designed to perform an in-depth analysis of trending hashtags and users. First, the module focuses on analyzing the content of hashtags by calculating the distribution of local trends, natural language, and hashtag reach. The Twitter trending panel often consists of hashtags related to local and global topics. Hashtags related to global topics like #MUFC and #CristianoRonaldo are discussed all around the world and these trends are not specifically related to local topics. We limit the scope of Manipify to study different aspects of local trending hashtags due to two major observations. First, global hashtags contain tweets in multiple languages. Second, these hashtags are adopted internationally and the context of hashtags varies with the cultural context of users[47]. These features of global hashtags made the understanding and classification of these hashtags a challenging task. Therefore, we propose a classification model of PK-Hash-Local to identify local trends for analysis. For classification, the following meta-information related to hashtags is fetched: (1) the trending time of hashtag in the target country, (2) trending time of hashtag in other countries, (3) list of countries hashtag seen in trending panel, and (4) whether hashtag has trended worldwide. We use this meta-information of hashtags to derive three features for the classification. The feature of "1st trend" is used to determine the country where a hashtag was first trended. The hashtag is likely to be related to a local topic if it is first seen in the trending panel of the target country. Similarly, the feature of "number of countries trended" counts the number of locations where the hashtag is seen in the trending panel. With a higher value of this feature, the hashtag has less probability of being a local trend. Finally, the self-descriptive binary feature of "worldwide trended" shows that the hashtag is global or local. Finally, we use meta-information features of hashtags to train the decision tree classifier.

To study the natural language of tweets related to a hashtag, we utilize the meta-information of tweet language provided by Twitter to calculate the language distribution as described in Section 3. Finally, the analyzer calculates the reach of a hashtag which is a measure to calculate the number of users who potentially have viewed the hashtag[§]. It also gives an approximation of the extent to which a trending hashtag can affect the Twitter community.

Furthermore, the trend analyzer performs the analysis of the user by studying the behaviour of bots and manipulating users. For user analysis, first, we detect manipulators and bots to investigate their characteristics. Next, their distribution is analysed in different hashtag categories. Besides, a time series analysis is presented for a sample of hashtags to scrutinize the behaviour of bots, humans, manipulators, and organic users.

5 Manipify Evaluation

In this section, first, we describe the performances of manipulators and bot classifiers. Next, we dive into the detail ed results of hashtag classification. Finally, we discuss the PK-Hash-Local classifier used to create a dataset for hashtags related to Pakistan only.

5.1 Manipulator detection

Table 5 presents the results of manipulator detection on MT-Dat. Due to brevity, we report the performance of the three best-performing classifiers only. We notice the competitive performance of LR and RF classifiers with 0.92 accuracy and F1-score value. Interestingly, the LR classifier shows superior performance in detecting manipulator accounts with a precision of 0.97 compared to 0.96 value of RF classifiers. For in-depth examination, we dive into the analysis of weights assigned to features by the best-performing LR classifier. We notice the highest contribution of the $\text{Tweets}_{\text{before}}$ feature in classification of manipulators. This result is expected as manipulators exhibit hyperactivity before the hashtag is first seen in the trending panel. In addition, the second-highest weight is assigned to the $\text{Sim}_{\text{score}}$ feature as the higher similarity between tweets of a user shows manipulative behaviour[40]. On the other hand, the most informative features to classify organic users are $\text{Time}_{\text{after}}$ and $\text{Time}_{\text{before}}$. In general, manipulators tend to post tweets with higher frequency to increase the volume and velocity of tweets[28]. Whereas, the large difference in time between consecutive tweets

[§]https: //www.tweetbinder.com/blog/twitter-impressions

Table 5 Classification performance on MT-Dat.

Classifier	Category	Precision	Recall	F1-score	Accuracy	AUC
LR	Manipulator	0.97	0.88	0.92	0.92	0.92
	Organic	0.88	0.97	0.92	0.92	0.92
	Overall	0.92	0.92	0.92	0.92	0.92
RF	Manipulator	0.96	0.89	0.92	0.92	0.92
	Organic	0.89	0.96	0.92	0.92	0.92
	Overall	0.92	0.92	0.92	0.92	0.92
GNB	Manipulator	0.94	0.84	0.89	0.89	0.89
	Organic	0.84	0.94	0.89	0.89	0.89
	Overall	0.89	0.89	0.89	0.89	0.89

assists in identifying organic users.

5.2 BotCat

Tables 6 and 7 shows the classification performance of the top three classifiers with the BT-Dat dataset. Focusing on Efthimion et al.'s[42] features, we note that the RF classifier achieves 0.87 accuracy on the BT-Dat dataset with the F1-score value of 0.87. Also, RF classifiers show the highest precision of 0.94 for the classification of bot accounts. Interestingly, the precision values of the top three classifiers for bot classification are in the range of 0.93–0.94 while the recall value of these classifiers is 0.81. This result signifies that the classifier assigns the label of the bot category with high confidence. In addition, bot users with less probability scores are assigned the label of human account. We also analyze the weights assigned by the RF classifier to the nine binary classification features. We observe that the classifier assigns the highest weight to the feature of follower count < 30 to the bot class explaining that users with a follower count less than 30 are more likely bot accounts. Similarly, the feature of user verified is assigned the highest weights for human accounts. In addition, we analyze bot accounts and observe that a few bot accounts have fake names, profile pictures, and the given description. Moreover, these accounts replicate the behaviour of human users which results in the misclassification of these users as human accounts[48].

Table 6 Classification results of Efthimion et al.[42] for bot classification.

Classifier	Category	Precision	Recall	F1-score	Accuracy	AUC
RF	Bot	0.94	0.81	0.87	0.87	0.87
	Human	0.81	0.94	0.87	0.87	0.87
	Overall	0.87	0.87	0.87	0.87	0.87
DT	Bot	0.93	0.81	0.87	0.86	0.87
	Human	0.81	0.93	0.86	0.86	0.87
	Overall	0.87	0.87	0.86	0.86	0.87
LR	Bot	0.93	0.81	0.87	0.87	0.86
	Human	0.81	0.93	0.86	0.87	0.86
	Overall	0.87	0.87	0.86	0.87	0.86

Table 7 Classification results of Yang et al.[43] for bot classification.

Classifier	Category	Precision	Recall	F1-score	Accuracy	AUC
DT	Bot	0.98	0.98	0.98	0.98	0.98
	Human	0.98	0.98	0.98	0.98	0.98
	Overall	0.98	0.98	0.98	0.98	0.98
RF	Bot	0.98	0.97	0.98	0.98	0.98
	Human	0.98	0.98	0.98	0.98	0.98
	Overall	0.98	0.97	0.98	0.98	0.98
KNN	Bot	0.97	0.98	0.98	0.97	0.97
	Human	0.98	0.96	0.97	0.97	0.97
	Overall	0.97	0.97	0.97	0.97	0.97

Yang et al.'s[43] features achieve an overall accuracy of 0.98 with DT and RF classifiers. Moreover, the DT classifier shows cutting-edge performance with 0.98 precision for bot identification. We further investigate 19 features according to the weights assigned by the best-performing DT classifier. We note that favourite count, user age, following-to-follower ratio, and growth rate of followers are the most contributing features in discerning the bot accounts. In general, lower favourite counts denote the bot accounts[49]. Similarly, bot accounts exhibit shorter ages because Twitter actively removes such automated and bot accounts from the platform. Furthermore, bot accounts follow a large number of users with fewer followers. Also, these accounts have negligible follower growth rate[50]. Finally, the high precision and high recall of classifiers

show the higher relevance and significance of Yang et al.'s[43] features in classifying bot accounts.

5.3 Hashtag classification

We train and evaluate the hashtag classifiers for English and Urdu hashtags separately. Table 8 gives the detailed classification results. First, we discuss the performance of the classifier for English hashtags. The English hashtag classifier achieved an overall accuracy of 0.84 with a 0.70 F1-score. However, we observe the highest F1-score of 0.97 for the military hashtags. Similarly, the sports and religion hashtags have comparable F1-scores of 0.93 and 0.94. Whereas, the other category hashtags achieve the lowest F1-score of 0.66. Finally, the politics, campaign, and entertainment hashtags have 0.82, 0.76, and 0.89 F1-scores, respectively. The low F1-score of hashtags for some categories is due to the high lexical diversity in samples of these classes. In addition, the number of training samples also affects the classification performance of binary classifiers.

Next, we divert our attention towards the classification results of Urdu hashtags. Overall, the Urdu hashtag classifier attains 0.79 accuracy and 0.63 F1-score. Here, the religion hashtags are classified with the highest F1-score of 0.93. Similar to the English hashtags, the other hashtags are classified with the lowest F1-score of 0.58. The entertainment hashtags also achieve a low F1-score of 0.65. Moreover, the categories of politics, sports, campaign, and military attain F1-scores of 0.82, 0.85, 0.74, and 0.81, respectively. Overall, we observe lower F1-scores for the Urdu language binary classifiers as compared to the English classifiers. We attribute this result to the lower number of samples in the training data for Urdu hashtags.

The classification of English-Urdu hashtags with our framework presents an interesting challenge which classifier (English or Urdu) should be used for such hashtags? In this regard, first, we classify the hashtag with both English and Urdu classifiers using tweets of respective languages. Next, we compare the probabilities for each class assigned by both classifiers. Finally, we assign the label of the category with the highest probability assigned by either classifier. Using this, approach, overall, we observe the English-Urdu classifier achieves 0.84 accuracy and 0.79 F1-score. However, we notice the best performance for the sports hashtags with 0.90 F1-score. The religion and politics category hashtags have comparable performance with 0.89 F1-score each. Whereas, the campaign, entertainment, and military hashtags are classified with 0.78, 0.79, and 0.84 F1-scores, respectively. Consequently, the other category hashtags are classified with the lowest F1-score of 0.41. We notice that this approach achieves better performance compared to the Urdu language classifier.

5.4 PK-Trends-Local

Leveraging the PK-Hash-Local classifier, we create the PK-Trends-Local dataset containing the trends related to Pakistan. In order to classify the "global" and "local" trends, first, we create the labelled dataset by manually labelling the 193 hashtags of the PK-Jan-21 dataset into the local and global category. With manual classification, 141 (73%) hashtags are labelled as local while 52 (27%) are labelled as global hashtags. Next, the meta-information related to hashtags in the PK-Trends dataset is fetched from GetDayTrends. Using the hashtag features (explained in Section 4) of the labelled dataset and the split ratio of 70:30, the classifier achieves the accuracy of 0.97. Next, we classify trends of PK-Nov-20 and PK-Dec-20 using the trained

Table 8 Hashtag classification results on HT-Dat dataset.

Language	Measure	Politics	Sports	Religion	Campaign	Entertainment	Military	Other
English	Precision	0.82	0.93	0.96	0.78	0.90	0.97	0.78
	Recall	0.82	0.92	0.92	0.72	0.88	0.97	0.71
	F1-score	0.82	0.93	0.94	0.76	0.89	0.97	0.66
Urdu	Precision	0.83	0.91	0.95	0.78	0.80	0.89	0.63
	Recall	0.82	0.83	0.92	0.72	0.65	0.79	0.61
	F1-score	0.82	0.85	0.93	0.74	0.65	0.81	0.58
English-Urdu	Precision	0.88	0.92	0.94	0.85	0.76	0.83	0.38
	Recall	0.91	0.89	0.85	0.72	0.83	0.86	0.44
	F1-score	0.89	0.90	0.89	0.78	0.79	0.84	0.41

classifier and identify 231 and 161 local hashtags in PK-Nov-20 and PK-Dec-20, respectively. Table 4 shows the distribution of local and global trends in PK-Trends.

6 PK-Trends Analysis

In this section, first, we classify and analyse the content of hashtags in the PK-Trends-Local dataset. Next, we identify the malicious users for each hashtag and discuss their distribution in PK-Trends-Local. Finally, we present the category-wise analysis of users.

6.1 Content analysis

To begin with the analysis, first, we classify the hashtags in PK-Trends-Local. This is done in order to conduct the user analysis for various hashtag categories. Figure 2 shows the distribution of hashtags and tweets related to each category in the PK-Trends-Local dataset. The classification results show that 32%–40% hashtags belong to the politics category. This result highlights the interest of the general public in politics. In addition, 15%–32% of Twitter trending hashtags belong to the campaign category. Interestingly, 8%–40% tweets belonging to this category show the promotional efforts of users for campaign hashtags. Moreover, only 2%–11% hashtags belong to the sports, 8%–10% to the religion, 5%–15% to the entertainment, and 1%–7% to the military category. Also, the "other" category contains less than 15% hashtags in three datasets of PK-Trends-Local, showing the large coverage of Manipify for analysis trending panel with six pre-determined categories. Zooming into the detailed analysis, we manually analyze the trends and note that the distribution of hashtag categories is intermittent due to the influence of real-world events on the trending panel. For instance, on the anniversary of the shooting at Army Public School (APS) in Pakistan on 16 December, 9 (35%) hashtags related to military class are seen in the trending panel.

Next, we analyze the distribution of natural languages of tweets, hashtag reach, and sentiments for different hashtag categories. As Manipify processes the data for Urdu and English languages only, we inspect the ratio of tweets for English, Urdu, unknown, and other languages as described in Section 3. Figure 3 shows the percentage of tweets of each language belonging to seven categories in PK-Nov-20. We notice that 15%–75% tweets are posted in the English language. Also, the politics and religion categories contain 56% and 60% tweets in the Urdu language, respectively. On the other hand, sports category contains 77% while the entertainment category contains 50% English tweets. Moreover, the dataset contains 60%–80% tweets posted in English and Urdu language, showing that the Manipify framework effectively analyzes the predominant part of tweets related to the trending panel. The datasets of PK-Dec-20 and PK-Jan-21 show a similar pattern for language distribution. From these results, we conclude that the user prefers the local language Urdu to discuss topics related to politics and religion categories. In addition, sports and entertainment categories contain a high percentage of English tweets because such category hashtags are discussed by international users as well. Focusing on the reach of the hashtag, we observe that

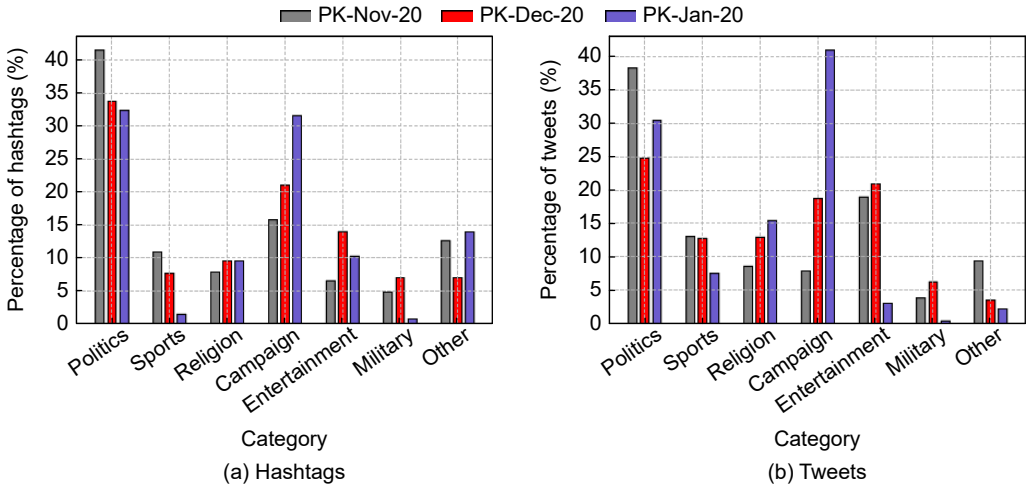

Fig. 2 Distribution of content for each category in PK-Trends-Local dataset.

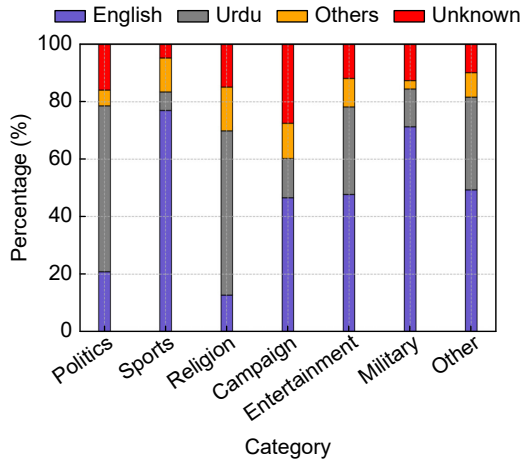

Fig. 3 Language distribution of tweets in PK-Nov-20.

the sports category has maximum reach with a limited number of hashtags and tweets (refer to Fig. 2). This result highlights that the substantial reach of the sports category is attributed to the usage of such hashtags by international celebrities. Moreover, the religion and campaign category hashtags have a lower value for reach. These results provide an interesting conclusion that religion category hashtags are generally used by normal Twitter users instead of celebrity users. Also, the campaign category hashtags are used for a limited audience from Pakistan.

6.2 User analysis

We initiate the user analysis by exploring the behaviour of users in PK-Trends-Local to determine the patterns of manipulation. Figure 4 shows the distribution of manipulators and organic users in PK-Trends-Local. Interestingly, PK-Nov-20, PK-Dec-20, and PK-Jan-21

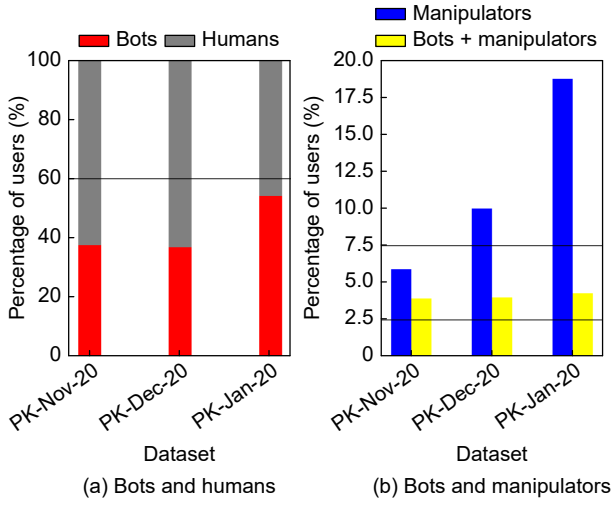

Fig. 4 Percentage of bots, humans, and manipulators.

contain 5.93%, 10%, and 18.7% manipulators, respectively. Focusing on bot accounts, PK-Nov-20, PK-Dec-20, and PK-Jan-21 contain 37%, 36%, and 54% bot accounts, respectively. As such, bot accounts are highly suspected to play a key role in the manipulation of the trending hashtags[51]. Therefore, we further investigate the users identified as bots as well as manipulators. Figure 4 shows the percentage of such users. We observe that 3.93%, 3.98%, and 4.3% accounts are identified as bots as well as manipulators in PK-Nov-20, PK-Dec-20, and PK-Jan-21, respectively. Interestingly, the percentage of bot users involved in manipulation is consistent in all data points. Whereas, the percentage of bots only is the least in PK-Nov-20 and the highest in PK-Jan-21. This result is expected because the accounts generating automated activity are suspended by Twitter[7]. Considering that the data related to all three datasets fetched in February 2021, the tweets posted by such deactivated or suspended accounts are not fetched in PK-Trends-Local.

Figure 5 shows category-wise distribution of manipulators, organic users, bots, and human accounts. Moreover, the percentage of tweets posted by these users is also provided. We observe the presence of only 1%–5% manipulating users in sports hashtags. This result is anticipated because sports hashtags like #PakvsSA are trended during the real-world event of a cricket match. However, a higher percentage of manipulators is observed in the politics and entertainment hashtags with 5%–10% and 8%–12% manipulators, respectively. On the contrary, the religion, campaign, and military categories contain only 2%–8% manipulators. However, focusing on the tweets, higher percentages of tweets are seen with a very low percentage of manipulative users. To sum up the results, we conclude that a higher percentage of manipulators in the politics hashtags is due to targeted mudslinging generated by the rival politics factions[52]. The manipulators generate fake trends of such politics hashtags to increase the exposure of their content to Twitter users. In addition, the trends of entertainment hashtags are generated with pre-planned coordinated efforts to promote TV shows, movies, and music. Shifting to the bot and human accounts, we notice that the campaign category has the largest percentage of bots with 60%–78% bots. Similarly, the sports hashtags contain 45%–50% bots. In addition, 42%–50% tweets related to politics hashtags are created by bot users.

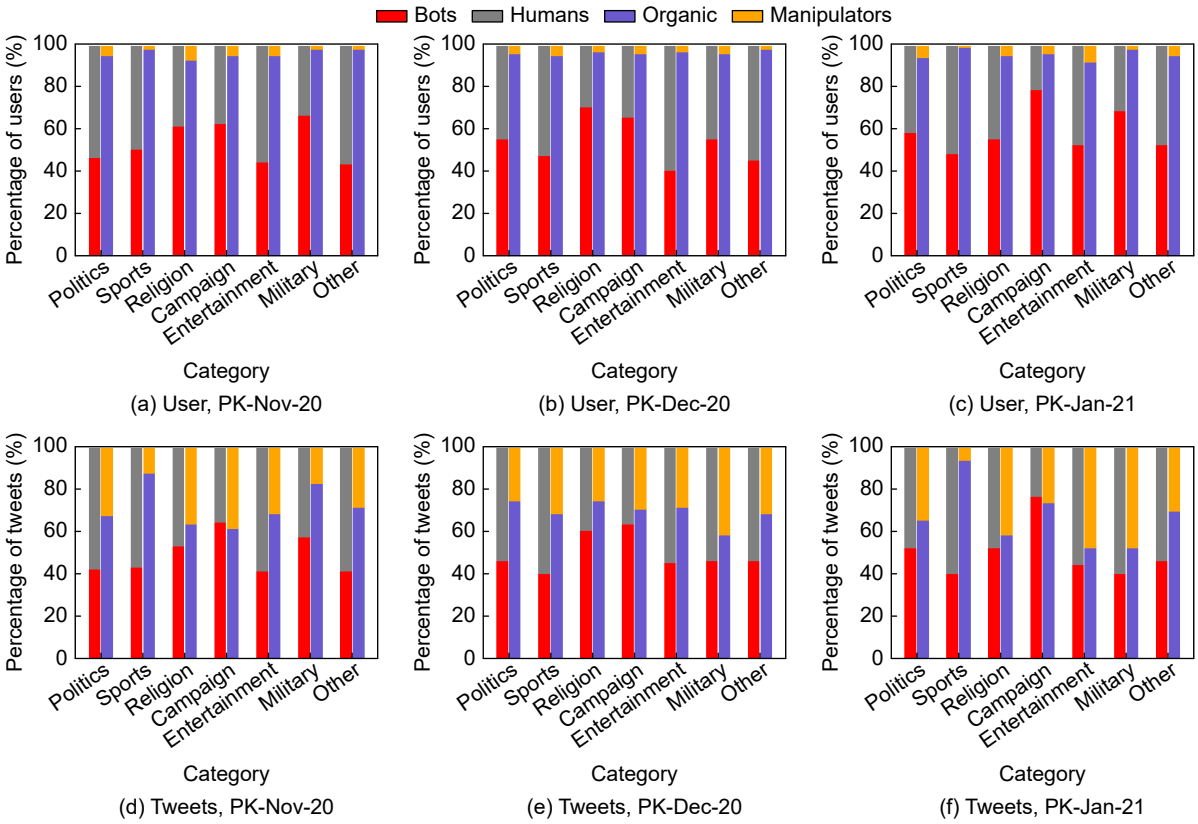

Fig. 5 Percentage of users and tweets of bots, humans, and manipulators in PK-Trends-Local dataset.

From these results, we conclude that bot accounts are used for the promotion of campaign and entertainment hashtags[53]. Moreover, sports hashtags like #PakvsSA are used by bot accounts to provide live updates related to the match. Interestingly, for politics hashtags, the bot accounts are used for promotion as well as to provide live updates related to political activity.

An interesting facet of user examination is the behaviour of users in different categories. For this purpose, first, we examine the dynamic evolution of manipulators and notice that 4.5×10^4 (80%) of accounts identified as manipulators are involved in the manipulation of only one hashtag. In addition, 12% users manipulate 2 trends while 8% users are identified in manipulating more than 2 hashtags. Zooming into such users manipulating more than one trend, interestingly, 763 users manipulate politics as well as entertainment category hashtags. Similarly, the religion and entertainment category has 253 overlapping manipulator accounts. In a similar vein, an inspection of bot accounts in different categories, we notice that politics and entertainment category hashtags have tweets from 4560 common bot accounts. In a similar vein, 3580 bot accounts posted tweets related to both entertainment and politics category hashtags. Next, we shift our intention to 916 human manipulators manipulating more than one hashtag. Interestingly, these users manipulate only one category hashtag. No human account involved in manipulation takes part in more than one category hashtag. This result is expected as human accounts focus on the manipulation of trends to spread their agenda/views. On the other hand, bot accounts are generally used by companies to create fake trends for economical/political benefits.

7 Conclusion

Manipulation of Twitter trends is a pervasive phenomenon. Automatic detection of manipulators is a challenging task because such users encompass the features of spam, bot, fake, and human accounts. In this paper, we propose a novel framework of Manipify to detect users manipulating the trending panel with 0.92 accuracy. We further identify bot accounts and classify trends into six categories. We inspect the users in the trending panel of Pakistan leveraging our framework with the PK-Trends dataset. We observe that the trending panel is dominated by politics category hashtags. In addition, politics and entertainment

category hashtags are the most manipulated trends in Pakistan. Moreover, 39% and 8% of users posting tweets related to trends in Pakistan are bots and manipulators, respectively. Furthermore, we find 8% manipulators generate 32.5% of tweets in the PK-Trends dataset. Finally, sports category hashtags have the least number of manipulators and maximum reach compared to other categories.

Manipify is a generic framework with language-independent features adaptable to examine manipulators disseminating content in different natural languages. In future, we plan to extend the framework to identify users who spread hate speech and propaganda in a coordinated manner. Besides, considering the multi-faceted 5th generation war on social media, we will work on location identification of manipulators working for rival countries to create polarization in society.

Acknowledgment

This work was supported by Higher Education Commission (HEC) Pakistan and Ministry of Planning Development and Reforms under National Center in Big Data and Cloud Computing.

References

[1] Twitter trends faqs, https://help.twitter.com/en/using-twitter/twitter-trending-faqs, 2021.

[2] A. Karami, L. S. Bennett, and X. He, Mining public opinion about economic issues: Twitter and the US presidential election, *Int. J. Strateg. Decis. Sci.*, vol. 9, no. 1, pp. 18–28, 2018.

[3] T. Rosenstiel, J. Sonderman, K. Loker, M. Ivancin, and N. Kjarval, Twitter and the news: How people use the social network to learn about the world, https://www.americanpressinstitute.org/publications/reports/survey-research/how-people-use-twitter-news/ , 2015.

[4] O. Gencoglu and M. Gruber, Causal modeling of Twitter activity during COVID-19, *Computation*, vol. 8, no. 4, p. 85, 2020.

[5] D. Assenmacher, L. Clever, J. S. Pohl, H. Trautmann, and C. Grimme, A two-phase framework for detecting manipulation campaigns in social media, in *Proc. International Conference on Human-Computer Interaction*, Copenhagen, Denmark, 2020, pp. 201–214.

[6] N. Abu-El-Rub and A. Mueen, BotCamp: Bot-driven interactions in social campaigns, in *Proc. World Wide Web Conference*, San Francisco, CA, USA, 2019, pp. 2529–2535.

[7] E. Gallagher, Manipulating trends & gaming twitter, https://erin-gallagher.medium.com/manipulating-trends-\\gaming-twitter-6fd31714c06c, 2016.

[8] B. Kollanyi, P. N. Howard, and S. C. Woolley, Bots and automation over Twitter during the US election, https://demtech.oii.ox.ac.uk/research/posts/bots-and-automation-over-twitter-during-the-u-s-election/, 2016.

[9] J. Uyheng and K. M. Carley, Bots and online hate during the COVID-19 pandemic: Case studies in the United States and the Philippines, *Journal of Computational Social Science*, vol. 3, no. 2, pp. 445–468, 2020.

[10] B. Nimmo, Measuring traffic manipulation on Twitter, https://demtech.oii.ox.ac.uk/research/posts/measuring-traffic-manipulation-on-twitter, 2019.

[11] T. Elmas, R. Overdorf, A. F. Özkalay, and K. Aberer, Ephemeral astroturfing attacks: The case of fake Twitter trends, in *Proc. 2021 IEEE European Symp. on Security and Privacy* (*EuroS&P*), Vienna, Austria, 2021, pp. 403–422.

[12] X. Dong and Y. Lian, A review of social media-based public opinion analyses: Challenges and recommendations, *Technol. Soc.*, vol. 67, p. 101724, 2021.

[13] S. Kausar, B. Tahir, and M. A. Mehmood, Push-to-trend: A novel framework to detect trend promoters in trending hashtags, *IEEE Access*, vol. 10, pp. 113005–113017, 2022.

[14] A. Zubiaga, D. Spina, V. Fresno, and R. Martínez, Classifying trending topics: A typology of conversation triggers on Twitter, in *Proc. 20th ACM Int. Conf. Information and Knowledge Management*, Glasgow, UK, 2011, pp. 2461–2464.

[15] Y. Zhang, X. Ruan, H. Wang, H. Wang, and S. He, Twitter trends manipulation: A first look inside the security of Twitter trending, *IEEE Trans. Inf. Forensics Secur.*, vol. 12, no. 1, pp. 144–156, 2016.

[16] H. U. Khan, S. Nasir, K. Nasim, D. Shabbir, and A. Mahmood, Twitter trends: A ranking algorithm analysis on real time data, *Expert Syst. Appl.*, vol. 164, p. 113990, 2021.

[17] R. Motamedi, S. Jamshidi, R. Rejaie, and W. Willinger, Examining the evolution of the Twitter elite network, *Soc. Netw. Anal. Min.*, vol. 10, no. 1, p. 1, 2020.

[18] Z. Wood-Doughty, M. Smith, D. Broniatowski, and M. Dredze, How does Twitter user behavior vary across demographic groups? in *Proc. Second Workshop on NLP and Computational Social Science*, Vancouver, Canada, 2017, pp. 83–89.

[19] U. Yaqub, S. A. Chun, V. Atluri, and J. Vaidya, Analysis of political discourse on Twitter in the context of the 2016 US presidential elections, *Gov. Inf. Q.*, vol. 34, no. 4, pp. 613–626, 2017.

[20] S. Kausar, B. Tahir, and M. A. Mehmood, Understanding the role of political micro-influencers in Pakistan, in *Proc. 2021 Int. Conf. Frontiers of Information Technology* (*FIT*), Islamabad, Pakistan, 2022, pp. 31–36.

[21] B. Tahir and M. A. Mehmood, Anbar: Collection and analysis of a large scale Urdu language Twitter corpus, *J. Intell. Fuzzy Syst.*, vol. 42, no. 5, pp. 4789–4800, 2022.

[22] V. Gupta and R. Hewett, Real-time tweet analytics using hybrid hashtags on Twitter big data streams, *Information*, vol. 11, no. 7, p. 341, 2020.
[23] D. M. Romero, B. Meeder, and J. Kleinberg, Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter, in *Proc. 20th Int. Conf. World Wide Web*, Hyderabad, India, 2011, pp. 695–704.
[24] M. Jeon, S. Jun, and E. Hwang, Hashtag recommendation based on user tweet and hashtag classification on Twitter, in *Proc. International Conference on Web-Age Information Management*, Macau, China, 2014, pp. 325–336.
[25] L. Posch, C. Wagner, P. Singer, and M. Strohmaier, Meaning as collective use: Predicting semantic hashtag categories on Twitter, in *Proc. 22nd Int. Conf. World Wide Web*, Rio de Janeiro, Brazil, 2013, pp. 621–628.
[26] P. Ferragina, F. Piccinno, and R. Santoro, On analyzing hashtags in Twitter, in *Proc. Ninth International AAAI Conference on Web and Social Media*, Oxford, UK, 2015, pp. 110–119.
[27] S. Kausar, B. Tahir, and M. A. Mehmood, HashCat: A novel approach for the topic classification of multilingual Twitter trends, in *Proc. 2021 Int. Conf. Frontiers of Information Technology* (*FIT*), Islamabad, Pakistan, 2022, pp. 212–217.
[28] S. Needle, How does Twitter decide what is trending? https://rethinkmedia.org/blog/how-does-twitter-decide-what-trending, 2016.
[29] Y. Hua, M. Naaman, and T. Ristenpart, Characterizing Twitter users who engage in adversarial interactions against political candidates, in *Proc. 2020 CHI Conf. Human Factors in Computing Systems*, Honolulu, HI, USA, 2020, pp. 1–13.
[30] Y. Hua, T. Ristenpart, and M. Naaman, Towards measuring adversarial Twitter interactions against candidates in the US midterm elections, *Proc. Int. AAAI Conf. Web Soc. Media*, vol. 14, pp. 272–282, 2020.
[31] K. -C. Yang, O. Varol, C. A. Davis, E. Ferrara, A. Flammini, and F. Menczer, Arming the public with artificial intelligence to counter social bots, *Hum. Behav. Emerg. Technol.*, vol. 1, no. 1, pp. 48–61, 2019.
[32] M. Mazza, S. Cresci, M. Avvenuti, W. Quattrociocchi, and M. Tesconi, RTbust: Exploiting temporal patterns for botnet detection on Twitter, in *Proc. 10th ACM Conf. Web Science*, Boston, MA, USA, 2019, pp. 183–192.
[33] K. Makice, *Twitter API: Up and Running: Learn How to Build Applications with the Twitter API*. Sebastopol, CA, USA: O'Reilly Media, 2009.
[34] T. Petersen, Mass scale manipulation of Twitter trends discovered, https://actu.epfl.ch/news/mass-scale-manipulation-of-twitter-trends-discov-2/, 2021.
[35] Digital 2023: Pakistan, https://datareportal.com/reports/digital-2023-pakistan, 2023.
[36] M. Mazhar, Fake tweets, bots and molding narratives: A look into Pakistani Twitter, https://tribune.com.pk/story/2343627/1, 2022.
[37] S. Popalzai and R. Jahangir, GrabYourKeyBoards: Inside Pakistan's hashtag mills, https://www.dawn.com/news/1519963, 2019.
[38] A. Baig, Misinformation warfare—#CivilWarinPakistan trends with 61% tweets coming from India; New Delhi contributes the highest number, https://digitalrightsmonitor.pk/misinformation-warfare-civilwarinpakistan-trends-with-61-tweets-coming-from-india-new-delhi-contributes-the-highest-number/, 2021.
[39] A. Gotter, Best time to post on Twitter in 2021? https://adespresso.com/blog/best-time-to-post-on-twitter/, 2021.
[40] Twitter's platform manipulation and spam policy, https://help.twitter.com/en/rules-and-policies/platform-manipulation, 2021.
[41] P. Dangeti, *Statistics for Machine Learning*. Birmingham, UK: Packt Publishing Ltd, 2017.
[42] P. G. Efthimion, S. Payne, and N. Proferes, Supervised machine learning bot detection techniques to identify social Twitter bots, *SMU Data Science Review*, vol. 1, no. 2, p. 5, 2018.
[43] K. -C. Yang, O. Varol, P. -M. Hui, and F. Menczer, Scalable and generalizable social bot detection through data selection, *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 1, pp. 1096–1103, 2020.
[44] F. Pérez-Hernández, S. Tabik, A. Lamas, R. Olmos, H. Fujita, and F. Herrera, Object detection binary classifiers methodology based on deep learning to identify small objects handled similarly: Application in video surveillance, *Knowl. Based Syst.*, vol. 194, p. 105590, 2020.
[45] T. Takenouchi and S. Ishii, Binary classifiers ensemble based on Bregman divergence for multi-class classification, *Neurocomputing*, vol. 273, pp. 424–434, 2018.
[46] K. Lee, D. Palsetia, R. Narayanan, M. M. A. Patwary, A. Agrawal, and A. Choudhary, Twitter trending topic classification, in *Proc. 2011 IEEE 11th Int. Conf. Data Mining Workshops*, Vancouver, Canada, 2012, pp. 251–258.
[47] P. Sheldon, E. Herzfeldt, and P. A. Rauschnabel, Culture and social media: The relationship between cultural values and hashtagging styles, *Behav. Inf. Technol.*, vol. 39, no. 7, pp. 758–770, 2020.
[48] S. Cresci, R. D. Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race, in *Proc. 26th Int. Conf. World Wide Web Companion*, Perth, Australia, 2017, pp. 963–972.
[49] R. Shukla, A. Sinha, and A. Chaudhary, TweezBot: An AI-driven online media bot identification algorithm for twitter social networks, *Electronics*, vol. 11, no. 5, p. 743, 2022.

[50] V. S. Subrahmanian, A. Azaria, S. Durst, V. Kagan, A. Galstyan, K. Lerman, L. Zhu, E. Ferrara, A. Flammini, and F. Menczer, The DARPA Twitter bot challenge, *Computer*, vol. 49, no. 6, pp. 38–46, 2016.

[51] F. Abdulrahman and A. Subedar, How much to fake a trend on Twitter?—BBC news, https://www.bbc.com/news/blogs-trending-43218939, 2018.

Soufia Kausar received the BSc degree in computer science from Lahore College for Women University, Lahore, Pakistan in 2017 and the MS degree in computer science from National University of Computing and Emerging Sciences (FAST-NU), Lahore, Pakistan in 2019. Currently, she is working as a research officer with the Al-Khawarizmi Institute of Computer Science (KICS), University of Engineering and Technology (UET), Lahore, Pakistan. Her research interests include machine learning, deep learning, computer vision, and text mining.

Muhammad Amir Mehmood received the PhD degree in engineering from Technische Universität Berlin/Deutsche Telekom Innovation Laboratories, Berlin, Germany in 2012. Currently, he has been working as an associate professor at the Al-Khawarizmi Institute of Computer Science, University of Engineering and Technology, Lahore, Pakistan. He has been the head of the High Performance Computing and Networking Lab (HPCNL) since 2013. His research interests include social computing, artificial intelligence, big data, and internet measurements.

[52] H. K. Evans, S. Smith, A. Gonzales, and K. Strouse, Mudslinging on Twitter during the 2014 election, *Soc. Media + Soc.*, doi: 10.1177/2056305117704408.

[53] Z. Gilani, L. Wang, J. Crowcroft, M. Almeida, and R. Farahbakhsh, Stweeler: A framework for Twitter bot analysis, in *Proc. 25th Int. Conference Companion on World Wide Web*, Montréal, Canada, 2016, pp. 37–38.

Bilal Tahir received the BSc degree in electrical engineering from National University of Computer and Emerging Sciences (FAST-NU), Lahore, Pakistan in 2014 and the MS degree in computer engineering from University of Engineering and Technology (UET), Lahore, Pakistan in 2018. Since 2017, he has been working as a senior research officer with the Al-Khawarizmi Institute of Computer Science (KICS), University of Engineering and Technology (UET), Lahore, Pakistan. His research interests include machine learning for images and text, natural language processing, deep learning, and information retrieval.