What to predict from Twitter Data?

Conference Paper · September 2023

DOI: 10.1109/ICCIT58132.2023.10273883

CITATIONS
2

READS
274

2 authors:

Mohammed Salemdeeb
Kocaeli University
13 PUBLICATIONS **43** CITATIONS
SEE PROFILE

Shaaban Sahmoud
Fatih Sultan Mehmet Waqf University
31 PUBLICATIONS **459** CITATIONS
SEE PROFILE

What to predict from Twitter data?

MOHAMMED SALEMDEEB
Electrical-Electronics Engineering
Istanbul Sabahattin Zaim University
Istanbul, Turkey
mohammed.salem@izu.edu.tr, ORCID: 0000-0002-2913-7671

SHAABAN SAHMOUD
Computer Engineering Department
Fatih Sultan Mehmet Vakıf University
Istanbul, Turkey
ssahmoud@fsm.edu.tr, ORCID: 0000-0003-0148-2382

***Abstract*— In the last decade, Twitter data has become one of the most valuable research sources in many areas such as health, marketing, security, and politics. Twitter data is preferred by researchers because it is completely public and can be easily downloaded using Twitter APIs. The recent intensive use of Twitter data makes it difficult to follow and analyze its research. In this paper, we summarize most of the predictable patterns, aspects, and attitudes from twitter data and analyze their performance and feasibility. Moreover, we list and describe the current popular Twitter datasets that are used in a variety of domains and applications. The current challenges and research gaps of these algorithms are discussed, and some recommendations and suggestions are given for future works for different domains and applications.**

Keywords*— *Twitter, prediction from Twitter, Twitter data analysis, Twitter datasets, Twitter features, social media.

I. INTRODUCTION

Twitter is a social media company that was created in March 2006 in the USA. After the achieved success in restricted networks, in November 2013, the company went public [1]. Since that date, the number of Twitter users dramatically increased to reach more than 330 million Monthly Active Users (MAU) in 2019. Although Twitter no longer reports MAUs after 2019, it is expected to have approximately 450 million MAU in 2022. This significant growth makes Twitter one of the world's most popular social media networks [2]. According to Statista's data as of January 2022, Twitter is in rank five in the topmost popular social media platforms for business after Facebook, YouTube, Instagram, and LinkedIn.

Twitter is defined as an online microblogging service that allows users to post short messages (no more than 280 characters) called tweets [3]. While the small number of allowed characters is one of the most important aspects of Twitter for a long time, recently there is a direction by the new Twitter CEO to increase it to 4000 characters in 2023. Although Twitter is mostly defined as a social media, many research papers and articles consider Twitter as an online news platform or sometimes both [4]. Moreover, Twitter gives the ability to their users to communicate and interact with each other through retweets, likes, replies, and quote tweets [5]. When a user types a tweet and posts it to Twitter's server, the other users (called followers) receive this message and become able to interact with it. Similarly, every user can see the tweets of their friends and can interact with them.

The Twitter platform has gained popularity as a source of data for research in a variety of fields. For example, Twitter data has been used in social media analysis where the data is used to study social interactions, opinions, and behaviours on the platform [6]. Researchers also have used Twitter data to study specific topics such as political polarisation, public sentiment, and the spread of information and misinformation [7]. In healthcare applications, Twitter data has been used to track the spread of diseases and to identify trends in drug usage [8]. In finance, researchers have used Twitter data to predict stock prices and trading volumes [9]. For marketing, many companies have used Twitter data to track consumer sentiment and to develop targeted marketing campaigns [10]. In most of these applications, Twitter has proven to be a valuable resource for research in a variety of fields due to its large-scale and real-time nature. One reason for the growing importance of Twitter is a large number of users on the platform, which has resulted in a vast amount of data being generated. This data can be used to study a wide range of phenomena and trends in real time, providing researchers with valuable insights that would not be possible to obtain through other platforms. In addition, Twitter has developed several application programming interfaces (API) and tools to make it easier for researchers to access and analyse Twitter data. These tools have made it more feasible for researchers to incorporate Twitter data into their studies, contributing to the growing importance of the platform in research [11].

In general, the features that can be directly extracted from Twitter platform are classified into four categories [12]. These categories are user features, time features, interaction features, and text features. The user features include general information about the considered Twitter account and its behaviour such as the number of followers, the number of friends, and the account age. The time features describe the time activities and habits of the Twitter account such as daily tweets rate and the average time between tweets. In interaction features, the interaction between Twitter users and their friends, their followers, and other users are used. Examples of features in this category include the number of likes for a tweet and the number of mentions for a Twitter user. The last category is called text features and it is used to characterise the text of tweets and replies such as counting hashtags, URLs, and emojis for a tweet [12].

In order to focus more on the increasing research interest on Twitter data, Fig. 1 shows the annual number of research articles on twitter published in IEEE and Scopus journals and conferences from 2009 to 2021. Twitter has started attracting the intention of IEEE community from 2009 where overall 9,365 research papers are found in IEEE xplore journals library until 2022. There are 12,887 research papers published in Scopus concerning twitter.

As shown in Fig. 1, the interest in twitter, as a source of data for the research community, has increased due to the modern prediction approaches and the developed technological tools such as GPUs and twitter APIs. It is expected to get more interest as it provides useful predicted information. However, there is a research gap between field-specific and survey research where the number of survey research in both Scopus and IEEE is 179 articles, and those articles are field-specific surveys; for example, a survey is conducted only on sentiment analysis using twitter dataset [13]. For predictions using Twitter, it is necessary to preprocess or clean the data to remove irrelevant information.

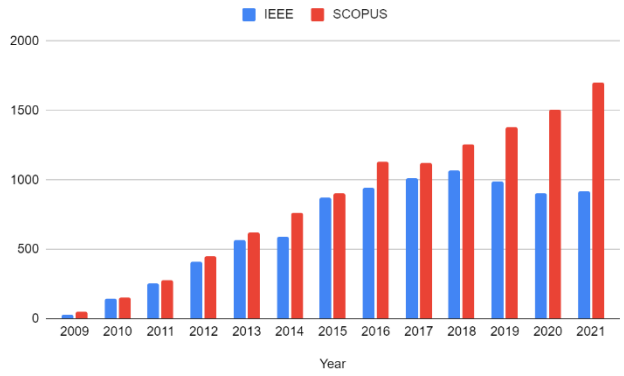

Fig. 1. The annual research papers on Twitter on IEEE and SCOPUS.

This may involve removing stop words, stemming or lemmatizing words, and handling missing or incomplete data. Next, relevant features must be selected and engineered to train a predictive model. There are a variety of machine learning (ML) algorithms that have been applied to Twitter data for prediction tasks including linear and logistic regression (LR), support vector machines (SVM), decision trees, and artificial neural networks (ANN) [14]. The choice of algorithm depends on the nature of the prediction task and the characteristics of the data. It is also important to consider issues such as overfitting and class imbalance when training predictive models on Twitter data [15].

In literature, there are many research articles published to predict a certain pattern from Twitter data but there is not enough effort spent to summarise these articles and methods. The contribution of this paper can be summarised as follows: (1) We summarise most of the predicted patterns, aspects, and attitudes from twitter data and analyse their performance and feasibility. (2) We list and describe the most important currently available Twitter datasets that are used in a variety of domains and applications. (3) We discuss the current challenges and research gaps of the prediction from Twitter algorithms. We also give some recommendations and suggest future works for different domains and applications. The organization of this paper is as follows: section one gives an introduction about Twitter and the prediction opportunities on Twitter. Section 2 describes the public and custom datasets extracted from Twitter and how to use them in research. Section 3 summarises the research categories and approaches over Twitter data. Section 4 concludes the paper and gives some recommendations for future work.

II. Datasets

Dataset preparation or data collection is one of the most challenging steps for analysing the social media data. The reason for this is the strictly applied laws associated with the use of public data and its usage as in Facebook and Twitter. Therefore, many researchers prefer to gather their datasets directly from social media platforms. Fortunately, Twitter has one of the fastest and easiest APIs to help developers and researchers access and gather data. Twitter API platform provides broad access to public Twitter data that users have chosen to share with the world. Using Twitter APIs, researchers can easily collect the targeted dataset after getting the required access keys and tokens. The Twitter API has a rich set of programmatic endpoints that allow researchers to retrieve and create various datasets, including tweets, users, and spaces [11]. Moreover, many libraries from different programming languages have used Twitter API to build flexible data-gathering tools like Tweepy, Twint, and Knime.

Usually, by using Twitter API, two main categories of data can be downloaded: User data and tweet data. Table I describes the information that can be retrieved for each category. As seen from the table, there are 12 fields/features for the tweets data category and 14 for the user data category. Most of these fields are either type string or integer except the user profile image. While these fields look simple and few, researchers can use them to generate hundreds of new features. For example, composing these fields with the four types of tweets (tweet, retweet, reply, and quote tweet) generates four different feature sets. Moreover, some of the fields as the tweet text contain a diverse set of characters, symbols, and links which can generate many features. As in many other domains, by analysing the existing publicly available Twitter datasets, it is found that most of them are available on data science and ML repositories such as Kaggle, GitHub, and UCI. Table II summarises a group of the most used Twitter datasets in research. As seen from the table, most of the datasets depend mainly on the tweet text to extract the features, especially for the sentiment analysis tasks. Obviously, Kaggle is the largest source of Twitter datasets since most of them are available publicly there. This is because Kaggle is a highly used community platform for data scientists and ML researchers. Moreover, it allows users to collaborate with other users, find and publish datasets, and use GPU-integrated notebooks for free.

TABLE I. The Downloadable Features using Twitter API

Category	# of features	Features
Tweets Data	12	Tweet text, tweet ID, time of the tweet, likes count, retweets count, is favorited, is retweeted, is it a retweet, retweet from, longitude, latitude, country.
User Data	14	User screen name, username, user ID, user profile image, user description, user URL, user creation time, user language, user location, user time zone, number of followers, number of friends, number of statuses, number of favourites.

TABLE II. The Most Used Twitter Datasets in Research

Dataset	Source	Task	# of Instances
Health News in Twitter [16]	UCI	Clustering	58,000
State-backed Trolls	Twitter	Classification	Very Large
AT-ODTSA [17]	GitHub	Sentiment Analysis	3,000
Sundanese Twitter [18]	UCI	Emotion Classification	2,510
ASTAD [19]	GitHub	Sentiment Analysis	36,000
Ara-SenTi-Tweet [20]	GitHub	Sentiment Analysis	17,573
Sentiment140	Kaggle	Sentiment Analysis	1.6 million
Twitter US Airline	Kaggle	Sentiment Analysis	14,579
User Gender	Kaggle	Classification	20,000
Customer Support	Twitter/ Kaggle	Natural Language Understanding	Very Large
Hateful Users [21]	Kaggle	Classification	100,000
Twitter Edge Nodes	Kaggle	Graph Analysis	11.3 million
Bot Datasets on Twitter [22]	-	Classification	Very Large
Psychopathy Prediction	Kaggle	Classification	2927
UtkML Spam Detection	Kaggle	Classification	784
Personality Prediction [23]	Kaggle	Prediction	2927
Covid-19 Twitter chatter	Kaggle	Prediction	Very Large
SMILE Twitter Emotion	Kaggle	Classification	3,085

III. CLASSIFICATION OF TWITTER IN RESEARCH

Usually, for social scientific research, the data are collected through traditional ways such as questionnaires and interviews which have limitations regarding efforts, time, volume, and reliability. Alternatively, the social media networks provide real information where specific data can be collected in a little time. Therefore, Twitter is used for a wide range of prediction tasks. In this paper, a sample of 133 research papers from SCI and SCOPUS-indexed journals, published from 2009 until 2020, were selected to figure out the scientific research interests regarding prediction from Twitter using artificial intelligence. The research interests are divided into 15 categories as shown in Fig. 2. It is found that the greatest research interest in Twitter is the public opinion and event prediction with 21.05% of research concentrated on this topic. The topic of least research interest is sports outcome prediction which has a percentage of 0.75% of the total number of the sample.

The second highest interested topic is the troll and spammer detection which obtain 17.29% of the research interest. In the third rank, 12.03% were interested in public health and disease prediction. 9.77% were interested in the user's location prediction. 8.27% were interested in stock price prediction. 7.52% were interested in election predictions. 5.26% were interested in link prediction and community detection. 4.51% were interested in personality and online behaviour prediction. 3.01% were interested in both crime prediction and gender/age prediction. 2.26% were interested in the user's interest area and also in security attack prediction. 1.5% were interested in sales prediction. Finally, 1.5% were interested in online ranking and popularity prediction. The most valuable research concerning Twitter is summarised in Table III. For each research, results, methodology and used datasets are presented. By analysing these research papers, we found that the most used ML techniques are SVM, LR, gaussian process regression, Ridge Regression (RR), k-Nearest Neighborhood (KNN), ANN and XGBoost. Moreover, the recent paradigms of deep learning mechanisms like Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) are starting to be used for predictions from Twitter datasets.

In line with the political nature of Twitter, political topics are one of the most studied and researched topics on Twitter. As shown in Table I, the election prediction topic has obtained significant interest and has been applied in many election cases such as in the USA, Pakistan, and France. Different algorithms were proposed to predict the results of these elections including decision trees and sentiment analysis methods [24-26]. The crime is also considered as a predictable behaviour from Twitter data. Many researchers utilised various classifiers such as LR and KNN to predict the probability of a crime occurring by using the criminals' shared tweets [27,28]. Predicting some human features such as gender and age is one of the earliest studied patterns from Twitter data [29]. The results of this research show a relatively very good performance with a percentage of more than 90%. The security attacks and its predictable patterns from twitter data are gaining increased attention from researchers. The research on this topic shows acceptable results where most of the machine learning techniques can be applied including Random Forest, XGBoost, and Naive Bayes [30, 31]. In link prediction and community detection, the research is done by first extracting the features from tweets, hashtags and links by ranking diagrams or node-attributed networks.

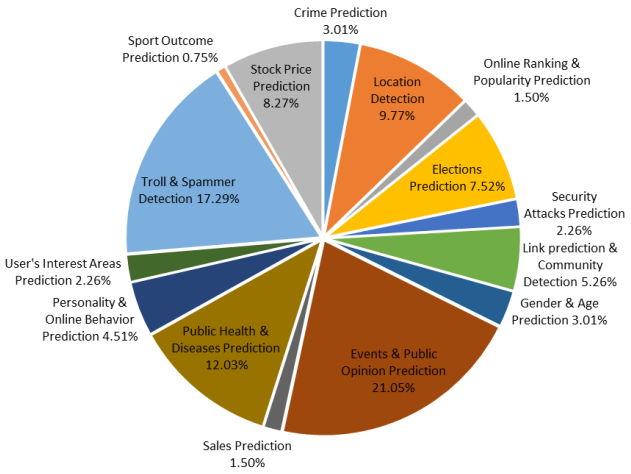

Fig. 2. Distribution of research interest in prediction from Twitter.

Then using a classifier to determine which community the tweet belongs to. The stochastic model classifier is the most used classifier that has acceptable accuracy of around 80% [32,33] for this type of problems. The Twitter user's personality aspects are also interested by many researchers. There are a lot of aspects that can be detected from Twitter data such as openness, extraversion, conscientiousness, agreeableness, and emotional stability. The user's tweets, status updates, timeline activities, and account information are analysed and classified to predict the personality aspect. Because many users do not share their location information or share fake locations, location prediction methods are widely studied by many researchers. The location information is considered significant information since it can be used in many applications such as advertisements and security domains. A Twitter user's location can be detected by utilising different approaches. Using the tweet GPS information and followers/friends' locations are two popular examples of these approaches. Usually, the location prediction methods obtain an acceptable accuracy between 70% and 80% because of the lack of shared information regarding the user's locations since many users prefer not to share their location data [37]. In [36] the authors used twitter users' gender, age, educational background, political stand, and personality to estimate the real location of that user. In addition, Twitter is used in E-commerce to introduce some predicting models for stock prices, Bitcoin exchange rate, interest rate and real estate market prices. Researchers stated that the number of tweets on Twitter can notably predict the Bitcoin trading volume and the future return which is important information for investors. Recently, many deep learning models based on CNNs were proposed to classify the sentiment of tweets and use it in stock price detection [38-40].

On the other side, researchers proved that Twitter can be used to predict diseases. Specifically, the users' tweets are used to train some deep learning models to detect the considered diseases such as influenza and depression. These models achieve acceptable results with a prediction accuracy around 90% [41-42]. Troll accounts and spammer users on Twitter are among the most common problems of the Twitter community. They cause public unrest and try to affect the public opinion by presenting and spreading unreal information and hate speech. For this reason, this topic attracts the research community to present new methodologies for predicting and filtering troll accounts.

TABLE III. OVERVIEW OF THE MOST VALUABLE RESEARCHES ON PREDICTION FROM TWITTER

Topic	Ref.	Methodology	Precision	Dataset	Properties
Elections Prediction	[24]	Decision Tree	99%	55k tweets	Pakistan
	[25]	AFINN online program	97.4%	7.541M tweets	USA
	[26]	Popularity Algorithm	98%	100k tweets	France
Crime Prediction	[27]	LR	67%	1.07M tweets	USA
	[28]	Naïve Bayes, KNN, SVM	94%	150k tweets	Cyber crimes
Gender & Age Prediction	[29]	LR	86%	3k users	Online TweetGenie Game
Security Attacks	[30]	ANN	96.73%	25,599* Instances	from early 2016 to March 2017
	[31]	FEEU for classification FRET for regression	80%	633k tweets, 51214 authors	XGBoost, SVM, LR, Naive Bayes
Link Prediction & Community Detection	[32]	Co-occurrence language networks	75%	39,882 Tweets	Introduce ranking diagram
	[33]	WTFW Stochastic Topic Model	81.2%	1.7M directed links	Nodes-attributed graphs
Personality & Online Behavior	[34]	SVM and XGBoost	97.99%	359 Twitter users	Indonesia language
	[35]	Gaussian Process Ridge Regression	92%	1.3K Twitter users	25 Tweets to Know You
Location Detection	[36]	LR	87%	4k users	user profiles
	[37]	LR	72%	1.3M users	accuracy in 161 km
Stock Price Prediction	[38]	CNN	0.0033 MAE**	9.6M Tweets	LR, SVM, and CNN
	[39]	Vector Autoregressive Model	------	Tweeted 'Bitcoin' 2014 to 2018	Predict Bitcoin's future return
	[40]	Residualized Control Approach	Pearson r =0.59	131M tweets	------
Public Health & Diseases	[41]	Specific Partial Differential Equation Model	90%	182k tweets over 18 weeks	Influenza prediction in USA
	[42]	CNN	87.43%	742,793 tweets	327 depressed users
Troll & Spammer Detection	[43]	Hawkes Processes Statistical Model	94.44%	1.8M images+ 9M tweets	3.6K Russian accounts
	[44]	Hawkes Processes Statistical Model	97.67%	10M posts by 5.5K Twitter & Reddit users	Russian & Iranian trolls
Events & Public Opinion Prediction	[45]	LSTM Stacked Autoencoder	93.45%	9275 tweets	Traffic network in the UK
	[46]	LR with Ordinary Least Squares	83%	16M tweets	Mentioning Congress Members

* Vulnerability instances
**The value is the average of mean average error (MAE)s of six events with sentiments.

Many machine learning techniques and frameworks such as Hawkes process statistical models are widely used to accomplish this task. Users' tweets, account information and sometimes the posted images are the most used features for troll accounts detection. Due to the large amount of research on this domain, the prediction accuracy is high with more than 95% [43,44].

Twitter is also used to detect some events and predict public opinion about a specific target. As an example, autoencoder and LR models are used to predict the traffic properties in the UK with a precision of 93.45% [45]. Twitter is also considered as a tool for users and decision-makers to express their opinions and for policymakers to get feedback that helps in taking decisions and to predict the acceptability, trust and transparency of those decisions. In early 2011, in the Middle East and during the Arab Spring, Twitter users played an important role in planning and organizing a sequence of protests. Therefore, much research has been successfully conducted to detect the tension and protests from Twitter data. Autoencoders, as a modern deep learning model, are mainly used in this prediction task by mentioning members of congress in the USA [46].

IV. CONCLUSION AND FUTURE DIRECTIONS

By utilising Twitter data, a variety of prediction tasks can be developed, analysed, and researched. The new machine learning models play a significant role in designing end-to-end frameworks for these prediction problems. Therefore, the number of research and studies that use Twitter data is expected to double in the next few years. After analysing, the research done using Twitter data, we found that some domains take more attention than others and there is still a lack of research noted for some domains. Moreover, from our comprehensive analysis, we discovered some issues and challenges that need more research work to be handled and solved. We summarize them in the following points: One of the most important problems that researchers face is the big amount of data needed for training and testing, especially for deep learning models. Gathering the data is not a big issue, but the quality of the data is the main concern. We found that many Twitter datasets have a low quality annotation which significantly degrades the performance of the trained models. As a result, there is an urgent need to prepare large datasets with reliable annotations. The link prediction and community detection topics need more research to test the new feature extraction methods and classifiers such as deep learning approaches. More research could be executed to predict the future sales volume of some products like iPhones, games and some applications. In addition, Twitter may be used to estimate disaster destruction such as earthquakes, floods and fires. Detecting the quality and people's feedback for some public services and institutions may also be considered as an interesting research problem. Most troll detection research deals with all types of troll accounts as one type which decreases the performance of the proposed algorithms. There is more research required to classify the types of troll accounts such as state-backed, bots and advertising troll accounts. The definition and properties for some prediction problems is not clear or misunderstanding. As an example, the location prediction is considered in many research as a nationality prediction problem as (since) some researchers try to detect the origin country of the users. There are more psychological, mental, social and health problems that can be considered to be predicted and discovered from Twitter data. The majority of current research uses datasets from Twitter only, where it is expected to be very beneficial and helpful to gather a dataset from different social media networks. One reason for this is the different characteristics of each social media network. As an example, studying a psychological health problem of a user by employing all his social media data will definitely enhance the results and can fully reflect his state of health.

REFERENCES

[1] I. Liu, C. Cheung, and M. Lee, "Understanding Twitter usage: What drive people continue to tweet," Pacific Asia Conference on Information Systems, PACIS 2010, Taipei, Taiwan, 2010.

[2] A. Smith and J. Brenner, "Twitter use 2012," Pew internet & American life project, pp 1-12, 2012.

[3] D. Murthy, Twitter: Social Communications in the Twitter age. Cambridge: Polity Press, 2018.

[4] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?," Proc. of the 19th inter. conf. on World wide web, 2010.

[5] S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts, "Who says what to whom on Twitter," Proceedings of the 20th international conference on World wide web, 2011.

[6] U. Kursuncu, M. Gaur, U. Lokala, K. Thirunarayan, A. Sheth, and I. B. Arpinar, "Predictive analysis on Twitter: Techniques and applications," Lecture Notes in Social Networks, pp. 67–104, 2018.

[7] M. Conover, J. Ratkiewicz, M. Francisco, B. Goncalves, F. Menczer, and A. Flammini, "Political polarization on Twitter," Proc. of the Inter. AAAI Conf. on Web and Social Media, vol. 5, no. 1, pp. 89–96, 2021.

[8] A. Signorini, A. M. Segre, and P. M. Polgreen, "The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic," PLoS ONE, vol. 6, no. 5, 2011.

[9] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," Journal of Computational Science, vol. 2, no. 1, pp. 1–8, 2011.

[10] J. Bollen, H. Mao, and A. Pepe, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena," Proc. of the Inter. AAAI Conf. on Web and Social Media, vol. 5, no. 1, pp. 450–453, 2021.

[11] S. Sahmoud and H. Safi, "Detecting Suspicious Activities of Digital Trolls During the Political Crisis," 2020 IEEE Inter. Conf. on Informatics, IoT, and Enabling Technologies (ICIoT), Doha, Qatar, 2020, pp. 532-537.

[12] S. Sahmoud, A. Abdellatif, and Y. Ragheb, "A fast algorithm for hunting state-backed Twitter trolls," Pervasive Computing and Social Networking, pp. 643–657, 2022.

[13] R. Wagh and P. Punde, "Survey on Sentiment Analysis using Twitter Dataset," 2018 Second Inter. Conf. on Electronics, Communication and Aerospace Technology (ICECA), 2018, pp. 208-211.

[14] Y. Zhang, X. Ruan, H. Wang, H. Wang and S. He, "Twitter Trends Manipulation: A First Look Inside the Security of Twitter Trending," in IEEE Transactions on Information Forensics and Security, vol. 12, no. 1, pp. 144-156, Jan. 2017, doi: 10.1109/TIFS.2016.2604226.

[15] M. Salemdeeb and S. Ertürk, "Full depth CNN classifier for handwritten and license plate characters recognition," PeerJ Comp. Sci., vol. 7, 2021.

[16] A. Karami, A. Gangopadhyay, B. Zhou, and H. Kharrazi, "Fuzzy approach topic discovery in health and medical corpora," International Journal of Fuzzy Systems, vol. 20, no. 4, pp. 1334–1345, 2017.

[17] S. Sahmoud, S. Abudalfa, and W. Elmasry, "At-ODTSA: A dataset of Arabic tweets for open domain targeted sentiment analysis," Inter. Journal of Computing and Digital Systems, vol. 11, no. 1, pp. 1299–1307, 2022.

[18] O. V. Putra, F. M. Wasmanson, T. Harmini and S. N. Utama, "Sundanese Twitter Dataset for Emotion Classification," 2020 Inter. Conf. on Comp. Eng., Network, and Intelligent Multimedia (CENIM), 2020, pp. 391-395

[19] K. A. Kwaik, S. Chatzikyriakidis, S. Dobnik, M. Saad, and R. Johansson, "An arabic tweets sentiment analysis dataset (ATSAD) using distant supervision and self training," in Proc. of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, 2020, pp. 1–8.

[20] N. Al-Twairesh, H. Al-Khalifa, A. Al-Salman, and Y. Al-Ohali, "Arasenti-tweet: A corpus for arabic sentiment analysis of Saudi tweets," Procedia Computer Science, vol. 117, pp. 63–72, 2017.

[21] M. Ribeiro, P. Calais, Y. Santos, V. Almeida, and W. Meira Jr., "Characterizing and detecting hateful users on Twitter," Proc. of the Inter. AAAI Conf. on Web and Social Media, vol. 12, no. 1, 2018.

[22] L. D. Samper-Escalante, O. Loyola-González, R. Monroy, and M. A. Medina-Pérez, "Bot datasets on Twitter: Analysis and challenges," Applied Sciences, vol. 11, no. 9, p. 4105, 2021.

[23] C. Sumner, A. Byers, R. Boochever, and G. J. Park, "Predicting dark triad personality traits from Twitter usage and a linguistic analysis of Tweets," 2012 11th Inter. Conf. on Machine Learning and Applications, 2012.

[24] T. Iqbal, F. Amin, W. Lohanna, A. Mustafa, and E. Sciences, "Mining Twitter Big Data to Predict 2013 Pakistan Election Winner," in IEEE International Conference on Multi Topic, 2013, pp. 49–54.

[25] M. Choy, M. Cheong, M. N. Laik, K. P. Shung, "US presidential election 2012 prediction using census corrected Twitter model," arXiv preprint arXiv:1211.0938, 2012.

[26] L. Wang and J. Q. Gan, "Prediction of the 2017 French election based on Twitter data analysis," 2017 9th Computer Science and Electronic Engineering (CEEC), Colchester, UK, 2017, pp. 89-93, doi: 10.1109/CEEC.2017.8101605.

[27] X. Chen, Y. Cho and S. Y. Jang, "Crime prediction using Twitter sentiment and weather," 2015 Systems and Information Engineering Design Symposium, Charlottesville, VA, USA, 2015, pp. 63-68.

[28] Z. Abbass, Z. Ali, M. Ali, B. Akbar and A. Saleem, "A Framework to Predict Social Crime through Twitter Tweets By Using Machine Learning," IEEE 14th Inter. Conf. on Semantic Computing, USA, 2020, pp. 363-368.

[29] D. Nguyen, D. Trieschnigg, A. Dogruöz, R. Gravel, M. Theune, T. Meder, F. de Jong, "Why Gender and Age Prediction from Tweets is Hard: Lessons from a Crowdsourcing Experiment", In Proc. of COLING 2014, the 25th Conference on Computational Linguistics, pp. 1950-1961, Doblin 2014.

[30] A. Subroto and A. Apriyana, "Cyber risk prediction through social media big data analytics and Statistical Machine Learning," Journal of Big Data, vol. 6, no. 1, 2019.

[31] H. Chen, R. Liu, N. Park, and V. S. Subrahmanian, "Using Twitter to predict when vulnerabilities will be exploited," Proc. of the 25th ACM SIGKDD Inter. Conf. on Knowledge Discovery; Data Mining, 2019.

[32] S. Martinčić-Ipšić, E. Močibob, and M. Perc, "Link prediction on Twitter," PLOS ONE, vol. 12, no. 7, 2017.

[33] N. Barbieri, F. Bonchi, and G. Manco, "Who to follow and why," Proc. of the 20th ACM SIGKDD inter. conf. on Know. Disc. and data mining, 2014.

[34] D. Suhartono, V. Ong, A. D. Rahmanto, N. given Williem, A. E. Nugroho, E. W. Andangsari, and M. N. Suprayogi, "Personality prediction based on Twitter information in Bahasa Indonesia," Proc. of the 2017 Federated Conference on Computer Science and Information Systems, 2017.

[35] P.-H. Arnoux, A. Xu, N. Boyette, J. Mahmud, R. Akkiraju, and V. Sinha, "25 tweets to know you: A new model to predict personality with social media," Proc. of the Inter. AAAI Conf. on Web and Social Media, vol. 11, no. 1, pp. 472–475, 2017.

[36] S. Volkova, Y. Bachrach, and B. Durme, "Mining user interests to predict perceived psycho-demographic traits on Twitter," 2016 IEEE Sec. Inter. Conf. on Big Data Computing Service & App. (BigDataService), 2016.

[37] A. Rahimi, T. Cohn, and T. Baldwin, "Twitter user geolocation using a unified text and network prediction model," Proc. of the 53rd Ann. Meeting of the Association for Computational Linguistics and the 7th Inter. Joint Conf. on Natural Language Processing (Volume 2: Short Papers), 2015.

[38] M. Yasir, S. Afzal, K. Latif, G. Chaudhary, N. Y. Malik, F. Shahzad, and O.-young Song, "An efficient deep learning based model to predict interest rate using Twitter sentiment," Sustainability, vol. 12, no. 4, p. 1660, 2020.

[39] D. Shen, A. Urquhart, and P. Wang, "Does Twitter predict bitcoin?," Economics Letters, vol. 174, pp. 118–122, 2019.

[40] M. Zamani and H. A. Schwartz, "Using Twitter language to predict the real estate market," Proc. of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, 2017.

[41] Y. Wang, K. Xu, Y. Kang, H. Wang, F. Wang, and A. Avram, "Regional influenza prediction with sampling Twitter data and PDE model," Inter. Jour. of Env. Research and Public Health, vol. 17, no. 3, p. 678, 2020.

[42] A. Husseini Orabi, P. Buddhitha, M. Husseini Orabi, and D. Inkpen, "Deep learning for depression detection of Twitter users," Proc. of the Fifth Work. on Comp. Ling. and Clinical Psyc.: From Keyboard to Clinic, 2018.

[43] S. Zannettou, T. Caulfield, B. Bradlyn, E. De Cristofaro, G. Stringhini, and J. Blackburn, "Characterizing the use of images in state-sponsored information warfare operations by Russian trolls on Twitter," Proc. of the Inter. AAAI Conf. on Web and Social Media, vol. 14, pp. 774–785, 2020.

[44] S. Zannettou, T. Caulfield, W. Setzer, M. Sirivianos, G. Stringhini, and J. Blackburn, "Who let the trolls out?," Proceedings of the 10th ACM Conference on Web Science, 2019.

[45] A. Essien, I. Petrounias, P. Sampaio, and S. Sampaio, "A deep-learning model for urban traffic flow prediction with traffic events mined from Twitter," World Wide Web, vol. 24, no. 4, pp. 1345–1368, 2020.

[46] Y. Theocharis, P. Barberá, Z. Fazekas, and S. A. Popa, "The dynamics of political incivility on Twitter," SAGE Open, vol. 10, no. 2, 2020.