SURVEY UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Real-Time Twitter Data Analysis: A Survey

Hakar Mohammed Rasul, Alaa Khalil Jumaa

Technical college of Informatics, Sulaimani Polytechnic University, Sulaimani 46001, Kurdistan Region, Iraq

ABSTRACT

Internet users are used to a steady stream of facts in the contemporary world. Numerous social media platforms, including Twitter, Facebook, and Quora, are plagued with spam accounts, posing a significant problem. These accounts are created to trick unwary real users into clicking on dangerous links or to continue publishing repetitious messages using automated software. This may significantly affect the user experiences on these websites. Effective methods for detecting certain types of spam have been intensively researched and developed. Effectively resolving this issue might be aided by doing sentiment analysis on these postings. Hence, this research provides a background study on Twitter data analysis, and surveys existing papers on Twitter sentiment analysis and fake account detection and classification. The investigation is restricted to the identification of social bots on the Twitter social media network. It examines the methodologies, classifiers, and detection accuracies of the several detection strategies now in use.

Index Terms: Twitter, Data Analysis, Twitter Streaming Application Programming Interface, Sentiment Analysis, Bot Detection

1. INTRODUCTION

The availability of data has increased dramatically since the Big Data era began, and it is predicted that this trend will continue in the years to come. A thorough research is being done to make appropriate use of this knowledge. Big data and big data analytics have created opportunities for businesses and scholars that were previously unthinkable. Research in artificial intelligence on how to leverage readily available data is producing fascinating and important results. There are several sources of big data, and one of the most well-known ones is social networks, including Twitter.

Twitter is a microblogging social network that enables users to post short messages (up to 280 characters) called tweets. Users may interact with one another on Twitter by responding to tweets, referencing other users in their tweets, or retweeting another user's message. Users can also follow each other to keep up with what other people are saying on Twitter. All registered users have access to the social network's services through a web page, mobile apps, and an application programming interface (API). The latter method of access has produced an ecosystem of applications that enhance the user's experience of information consumption and aggregation. However, this has aided in the development of systems for account management and automated tweet publishing [1]. Fully-automated accounts are called bots. They can retweet exciting and relevant material for specific communities or aggregate tweets about a topic.

One area that requires attention is bot detection analysis. Since, around 48 million Twitter accounts have been maintained by automated programs dubbed bots, accounting for up to 15% of all Twitter accounts [2]. Certain bots are helpful for a numerous task, including automatically publishing news and academic articles and aiding in emergency circumstances. Nonetheless, Twitter bots have been used for malicious

Access this article online

DOI:10.21928/uhdjst.v6n2y2022.pp147-155 | E-ISSN: 2521-4217 P-ISSN: 2521-4209

Copyright © 2022 Rasul and Jumaa. This is an open access article distributed under the Creative Commons Attribution Non-Commercial No Derivatives License 4.0 (CC BY-NC-ND 4.0)

Corresponding author's e-mail: Technical college of Informatics, Sulaimani Polytechnic University, Sulaimani 46001, Kurdistan Region, Iraq. E-mail: Hakar.mohammed.r@spu.edu.iq

Received: 17-06-2022 Accepted: 11-11-2022 Published: 21-12-2022

purposes, such as spreading malware or manipulating public opinion on a certain subject.

Bot identification software is predicated on the premise that the behavior of a human account is distinct from that of a bot. To quantify these discrepancies, representative factors including the statistical distribution of the terms used in tweets, the frequency of daily posts, and the number of individuals who followed the user may be employed [3]. Apache Spark data analysis on Twitter will be required to do that. As a result, the portion that follows in this essay will examine similar efforts on Twitter data analysis employing Apache Spark and bot identification, as well as the available tools.

2. BACKGROUND INFORMATION

2.1. Twitter

Twitter is a microblogging and social networking website that enables users to post and receive 280-character messages called "tweets." Registered users may send tweets and follow other users. Unregistered users may browse public tweets on Twitter without having an account [4].

Over 300 million individuals use Twitter on a regular basis. More than 500 million tweets each day are sent in 33 different languages [5]. One of Twitter's best benefits is the capacity for communication and sharing with other users. By sharing links, pictures, and videos with their followers, people and businesses may interact with them [6]. This section explains some of Twitter features:

1. Follow: To follow someone on Twitter, you must subscribe to their tweets or site updates. Another Twitter user who has followed you is referred to as a "Follower." Other Twitter users you've decided to follow on the platform are referred to as "following." [7]
2. @: In tweets, the @ symbol is used to identify usernames. The @ sign before a username (like @HakarRasul) creates a connection to that Twitter user's profile [8].
3. Reply: A tweet in response to a tweet from another person. To answer to a tweet, users often click the "reply" box or icon adjacent to it. @username is always the first character in a reply [9].
4. Retweet: The act of forwarding another user's tweet is denoted as "retweeting." In essence, you are sharing another user's tweet in your profile while properly acknowledging the message's original writer [10].
5. Mention: This term refers to tweets that contain a username. @replies are a type of mention as well [11].
6. Hashtag: The # symbol is used in tweets to denote topics or keywords. Hashtags are limited to letters and numbers (no punctuation). Other Twitter users may use a hashtag you tweet to search for it. Any Twitter user may generate a hashtag at any moment [12].
7. Direct Messages: These Tweets, sometimes referred to as direct messages or simply "messages," are confidential between the transmitter and recipient. When you start a tweet with "d username" to identify the recipient, the tweet becomes a direct message (DM). You should be following someone to send them a Direct Message [13].
8. Trends: A subject recognized by Twitter's algorithm as among the hottest subjects on the network right now [14].
9. Favorites: To add a Tweet to your favorites, click the yellow icon next to the tweet. Tweets you've favorite will stay in your list until you delete them [15].

2.2. Twitter Streaming API

The Twitter API now includes a Streaming API in addition to two separate REST APIs. The streaming API provides real-time access to Tweets that were sampled and filtered. The API is HTTP-based, with data accessible through GET, POST, and DELETE requests. The streaming API allows you to access subsets of public status descriptions, such as answers and mentions from public accounts, in near-real time. Protected users' status descriptions and direct messages are no longer viewable. The streaming API may filter status descriptions based on quality criteria, which are influenced, in addition to, by frequent and repeated status updates [16].

The API requires a valid Twitter account and employs simple HTTP authentication. Data may be obtained in both XML and the shorter JSON format. The parsing of JSON data got from the streaming API is straightforward: Each object is delivered on a separate line, with a carriage return at the conclusion [17].

Twitter streaming data allow every user to learn about what is going on in the globe at any given moment. The Twitter streaming API provides access to a huge quantity of tweets in real time [18].

A Python package called Twitter4j is available to access the streaming API and download Twitter data to analyze data from the Twitter API. This data has been filtered using a list of provided keywords. This research will use Apache Spark, a distributed data processing system with many workers, and master nodes. This cluster can manage millions of records and is scalable. Map reduction on Spark might be used to filter out the massive amount of data. For each tweet in the data, a JSON object will be included in the input file. On the Spark frame structure, this file will be uploaded. The mapper classifies all files in the directory according to the filter specified once

the Spark frame structure has been duplicated and distributed across several nodes. These cleaned tweets will go through data mining techniques, allowing for a one-to-one analysis of data that will be useful for making difficult judgments [19].

2.3. Analysis Process on Twitter Data

There are several steps to be performed to analyze Twitter data. Fig. 1 illustrates the phases of analyzing Twitter data.

2.3.1. Dataset collection

To gather real-time data, an application must be developed that uses the Twitter API to capture the information of people who recently tweeted about the issue and construct a user-based feature set data frame [21].

2.3.2. Processing tweets

This stage involves removing unnecessary material from tweets in the style of regular expressions [22].

2.3.3. Feature selection

Here, some of features should be considered, such as the user-based and content-based. These features have to be selected to enhance the detection and classification process [21].

2.3.4. Classification

In this step, the user must be checked in real-time whether it is a bot or human, which may be accomplished by training and testing the proposed model. The proposed model can be build using one of the machine learning algorithms. After applying the machine learning algorithm on a preprocessed, existed, and labeled dataset, a model can be created. Then, this model can be used to predict if the streamed Twitter that we got from Twitter is human or bot [22].

3. METHODOLOGY

This section will provide the clarification of the searching, filtering, and stages that were employed throughout this paper's research stage.

3.1. Research Sections

In Section I, questions like (What is big data, what is Twitter, and what is the connection between Twitter and big data?) has been answered. Then, Section II explained Twitter and its important elements; Twitter API and its use; and the phases of Twitter data processing. Section III gives a methodology about how this paper been organized and the methods that have been used to gathered information. Section IV provides a survey methodology and has been divided into two parts survey of articles about sentiment analysis and survey of articles about bot detection and classifications.

Fig. 1. Twitter data analysis process [20].

3.2. Search Query

This paper aims to summarize the current state of the real-time Twitter data analysis topic and discuss the findings presented in recent research papers. Hence, those keywords have been used.

("Twitter data") AND ("Real-Time OR "Bots") AND ("Sentiment analysis" OR "Bot Classification" OR "Data Extraction" OR "Preprocessing" OR "Text-mining" OR "web-Mining") AND ("Challenges" OR "Problems" OR "Patterns").

3.3. Selection of Sources

Google Scholar and Elsevier have been used for applying the search queries and the databases that have been considered were IEEEXplore Digital Library, SpringerLink Journal, Elsevier, and Science Direct.

3.3.1. Selection phases

Each article that has been chosen to be used in this paper has been gone through these processes:

The first phase of article selection is applying the search queries. Then, select only the articles have been published between 2016 and 2021. After that, the title of the research and the list of index terms had been considering to see if it includes the keyword "Twitter, Data Analysis."

The next step was reading the abstract and the conclusion of the paper, and selecting the paper according to its abstract and conclusion then the relatively of its body to them. Finally, the last phase was considering the journal's indexing and if they are peer reviewed or not.

4. SURVEY METHODOLOGY

When coming to Twitter data analysis, there are various types of analysis that might be done on the collected data such as sentiment analysis, tweets classification, and fake tweets detection. Hence, this survey will be categorized into two sections (A) sentiment analysis and (B) tweets classification and bot detection. Table 1 shows list the studies that have been surveyed in this section.

TABLE 1: List of studies that have been reviewed in Section IV

Title (s)	Author (s)	Technique (s)	Result (s)	Year
Sentiment Analysis				
"Sentiment analysis and classification of Indian farmers' protest using twitter data"	Ashwin Sanjay Neogi, Kirti Anilkumar Garg, Ram Krishn Mishra, Yogesh K Dwivedi	Bag of Words and TF-IDF	Bag of Words was more effective than TF-IDF.	2022
"An optimal deep learning-based LSTM for stock price prediction using twitter sentiment analysis"	T. Swathi, N. Kasiviswanath, A. Ananda Rao	TLBO-LSTM	Precision: 0.95, Recall 0.85, Accuracy: 0.94, F1-score 0.90	2022
"Twitter Sentiment Analysis during COVID-19 Outbreak"	Akash Dutt Dubey	NRC Emotion Lexicon	The majority of individuals around the globe are optimistic.	2020
"Detection of Fake Tweets Using Sentiment Analysis"	C. Monica, N. Nagarathna	Rule-based prediction	Accuracy: 0.97, F1-score: 0.73, Precision: 1.00, Recall: 0.97	2020
"Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers"	Gonzalo A.Ruz, Pablo A. Henríquez, Aldo Mascareño	Bayes factor	Accuracy: 0.85, Precision: 0.92, Recall: 0.77, F1-score: 0.82	2020
"Twitter Sentiment Analysis Based on Ordinal Regression"	Shihab Elbagir Saad, Jing Yang	Multinomial logistic regression (SoftMax), Support Vector Regression (SVR), Decision Trees (DTs), and Random Forest (RF)	Accuracy: 0.91, F1-score: 0.85 using Decision Tree.	2019
Classification and bot detection				
"The Rise of Social Bots," "Online human-bot interactions: Detection, estimation, and characterization," "Deep Neural Networks for Bot Detection," "Evolution of bot and human behavior during elections," "Measuring bot and human behavioral dynamics"	Emilio Ferrara *et al*.	Session features	Accuracy: 0.97	2016 2017 2018 2019 2020
"A deep learning model for Twitter spam detection"	Zulfikar Alom. Barbara Carminati, Elena Ferrari	Deep learning	Accuracy: 0.99, Recall: 0.98, F1-score: 0.93	2020
Classification and bot detection				
"Twitter Bot Detection Using Bidirectional Long Short-Term Memory Neural Networks and Word Embeddings"	Feng Wei, Uyen Trang Nguyen	Recurrent neural networks, specifically bidirectional Long Short-term Memory (BiLSTM)	Accuracy: 0.92, Precision: 1.00, Recall: 0.85, F1-score: 0.92	2019
"Social Network Polluting Contents Detection through Deep Learning Techniques"	Fabio Martinelli, Francesco Mercaldo, Antonella Santone	Combination of word embedding and deep learning	Precision: 0.79, Recall: 0.73, F1-score: 0.76	2019
"DeepScan: Exploiting Deep Learning for Malicious Account Detection in Location-Based Social Networks"	Qingyuan Gong, Yang Chen, Xinlei He, Zhou Zhuang, Tianyi Wang, Hong Huang, Xin Wang, Xiaoming Fu	long short-term memory (LSTM) neural network	Precision: 0.95, Recall: 0.97, F1-score: 0.96	2018
"Measuring bot and human behavioral dynamics"	Iacopo Pozzana, Emilio Ferrara	Extra Trees (ET), DT, Random Forests (RF), Adaptive Boosting (AB), and KNN	ET and RF had the greatest cross-validated average performance 0.86	2018
"Deep neural networks for bot detection"	Sneha Kudugunta, Emilio Ferrara	Deep neural network based on contextual long short-term memory (LSTM)	Accuracy: 0.96, Precision: 0.96, Recall: 0.96, F1-score: 0.96	2018
"Classification of Twitter Accounts into Automated Agents and Human Users"	Zafar Gilani, Ekaterina Kochmar, Jon Crowcroft	Random Forests classifier	Accuracy: 0.86, Precision: 0.85, Recall: 0.82, F1-score: 0.83	2017

(*Contd...*)

TABLE 1: (*Continued*)

Title (s)	Author (s)	Technique (s)	Result (s)	Year
"Detecting Automation of Twitter Accounts: Are You a Human, Bot, or Cyborg?"	Zi Chu, Steven Gianvecchio, Haining Wang, Sushil Jajodia	Bayesian classification	overall system accuracy: 96.0	2012
"Detecting Spam Bots in Online Social Networking Sites: A Machine Learning Approach"	Alex Hai Wang	Decision Tree (DT), Neural Network (NN), Support Vector Machines (SVM), Naive Bayesian (NB), and k-Nearest Neighbors, are used to detect spam bots (KNN)	Accuracy: 0.91, Precision: 0.91, Recall: 0.91, F1-score: 0.91 using NB	2010

4.1 Sentiment Analysis on Twitter

A computer finding the mood of a word, phrase, or tweet is quite challenging. To ascertain the polarity of the words and perform sentiment analysis, human participation is required. Since it is used to evaluate people's sentiments, views, and emotions, this form of analysis is sometimes referred to as "opinion mining." It is done by evaluating each word's attitude and classifying it as either positive, negative, or neutral. In addition, there are other ways to do sentiment analysis, including by employing a lexicon, machine learning, deep learning, or a combination of machine learning and lexicon-based approaches. In the lines that follow, recent studies on sentiment analysis on Twitter will be reviewed.

Neogi *et al.* [23] acquired data from the microblogging website Twitter on farmer protests to comprehend the global views shared by the public. They categorized and analyzed the attitudes based on over 20,000 tweets about the demonstration using algorithms. Using Bag of Words and TF-IDF for their investigation, they observed that Bag of Words performed better than TF-IDF. In addition, they used Naive Bayes, Decision Trees, Random Forests (RFs), and Support Vector Machines and found that RF provided the most accurate categorization. Given that millions of individuals shared their thoughts about the protests, one of the study's limitations is that they may have retrieved a rather high number of tweets. A greater quantity of tweets may have been useful in revealing a variety of emotions.

Using Twitter data, Swathi *et al.* [24] provide a novel teaching and learning-based optimization (TLBO) model with long short-term memory (LSTM)-based sentiment analysis for stock price prediction. Due to the short length and peculiar grammatical patterns of tweets, data pre-processing is required to eliminate irrelevant information and put it into a readable format. In addition, the LSTM model is used to categorize tweets into positive and negative opinions about stock values. They help explore the correlation between tweets and stock market values. The Adam optimizer is used to set the learning rate of the LSTM model to enhance its prediction performance. In addition, the TLBO model is used to properly adjust the output unit of the LSTM model. On Twitter data, experiments are conducted to improve the forecasting ability of the TLBO-LSTM model for stock prices. The experimental results of the TLBO-LSTM model outperform the state-of-the-art approaches in a variety of respects. The TLBO-LSTM model gave an excellent result, with a maximum accuracy of 95.33%, a recall of 85.28%, and an F-score of 90%. The TLBO-LSTM model outperformed the competition by attaining a superior accuracy of 94.73%.

Dubey [25] used Twitter Sentiment Analysis to ascertain how residents in different countries are coping with the COVID-19 outbreak. The research analyzed tweets from 12 different countries. These tweets were gathered between March 11, and March 31, 2020, and are associated to COVID-19 in some manner. The tweets were acquired, pre-processed, and then subjected to sentiment and text mining analysis. The study's findings show that, although the most people worldwide are optimistic and hopeful, there are instances of fear, sadness, and disdain around the globe. The study analyzed tweets from the selected nations using the NRC Emtoion lexicon. The NRC Lexicon of Word-Emotion Associations has 10,170 lexical units that examine not just positive and negative polarity, but also the eight emotions established by Plutchik. On average, 50,000 tweets were used in the study from each nation every 4 days. The collection was conducted using the R package RTweet. COVID-19, coronavirus, corona, stay home stay safe, and COVID-19 pandemic were the keywords used to gather the tweets. While collecting the tweets, the retweets and responses were filtered out to prevent repetition. When the whole database was in hand, data cleaning was done,

which included the removal of white spaces, punctuation, stop words, and the conversion of tweets to lower case. Following data cleansing, the tweets were analyzed using the NRC Emtoion lexicon using the get_nrc sentiment function. After scoring tweets on feelings and emotions, a corpus was built to generate a word cloud for each nation. However, a drawback of the study is that the NRC Emtoion language does not include sarcasm and irony as emotions.

In another study, Monica and Nagarathna [26] give users who have recently written about a certain topic a model that analyzes how they feel about it based on real-time data. They use this algorithm to create a sentiment score for each user based on content-based criteria to detect Twitter spam. The suggested method applies a custom rule-based algorithm for bot detection and compares it to a number of different algorithms such as MLP, decision tree, and RF to establish the model's effectiveness in detecting spam accounts. The Twitter API was used to collect real-time data for this investigation. The data extraction procedure includes extracting the characteristics required for the research, preprocessing, and sentiment analysis. Then, using the Fake Prediction Algorithm, MLP, Decision Tree, and RF, the data are categorized to determine how many of them are authentic and legitimate users. They resulted that the rule-based fake prediction system achieved the score of accuracy of 0.97, which was superior to the existing machine learning classifiers. The study has two major limitations. First, the group of users from which data had been collected is small. Second, English was the only language examined for analysis.

Using data from the 2010 Chile earthquake and the 2017 Catalan independence vote, Ruz *et al.* [27] examined five classifiers (one of which is a variation of the TAN model) and evaluated their effectiveness on two Twitter datasets. They are considering Bayesian network classifiers for sentiment analysis on two Spanish-language datasets: The 2010 Chilean earthquake and the 2017 Catalan independence vote. To automatically manage the amount of edges supported by training instances in the Bayesian network classifier, they employ a Bayes factor technique, resulting in networks that are more realistic. Given a significant number of training instances, the findings demonstrate the efficacy of the Bayes factor measure and its competitive prediction performance when compared to support vector machines and RFs. In addition, the generated networks enable the identification of word-to-word relationships, so providing valuable qualitative information for understanding the key characteristics of event dynamics from a historical and social perspective. Even though there are not enough training examples, the research achieves that the event dynamics may be understood using qualitative information from TAN and BBF TAN. Furthermore, the generated networks may be applied to convey a tale about the important event that was studied. However, this study may be enhanced by applying the Bayesian network classifier and grounded theory.

Along the same line, Saad and Yang [28] effort to undertake a complete Twitter sentiment analysis using machine learning techniques and ordinal regression. The suggested technique comprises pre-processing tweets and then generating a relevant feature using a feature extraction method. The scoring and balancing aspects come next, and they may be categorized in a number of different ways. The suggested system uses RF, multinomial logistic regression (SoftMax), decision trees (DTs), and support vector regression (SVR) methods for sentiment analysis categorization. This system's real implementation is dependent on a Twitter dataset made available through the NLTK corpus resources. According to experimental data, the proposed solution may reliably detect ordinal regression using machine learning methods. Furthermore, the results suggest that Decision Trees outperform all other algorithms in terms of delivering the best outcomes. The proposed system consists of four key components. The first module is data acquisition, which is the method of gathering labeled tweets for sentiment analysis; the second module is preprocessing, which is the method of converting and refining tweets into a data set that might easily be used for further analysis. The third module emphases the extraction of relevant features for classification model construction. Following that, the method for balancing and evaluating tweets is presented. The final module sorts tweets into high positive, moderate positive, neutral, moderate negative, and high negative categories using a quantity of machine learning classifiers. Based on the study results, SVR and RF have almost the same accuracy, which is superior to the multinomial logistic regression classifier. The decision tree, however, is the most accurate, with a score of 91.81%. Based on the findings of the trials, the suggested model can accurately detect ordinal regression in Twitter using machine learning methods.

4.2. Fake Account Detection and Classification

Twitter bots are software-controlled automated Twitter accounts, while they are taught to perform duties similar to those carried out by regular Twitter users, such as like tweets and following other users. Twitter bots can be applied for a number of beneficial reasons, including broadcasting critical material such as weather crises in real time, publishing useful content in bulk, and producing automated direct message responses. However, Twitter bots might be used for negative purposes such spreading fake news campaigns, spamming, compromising others' privacy, and sock-puppetry. The

following paragraphs will be a survey of resent researches on Twitter bot detection and classification.

Kudugunta and Ferrara [29] used both conventional machine learning classifiers and deep learning techniques to identify bots on Twitter, both at the account and tweet levels. They used SMOTE with data augmentation using (1) Edited Nearest Neighbors (ENN) and (2) Tomek Links to address the unbalanced dataset. A collection of classifiers, including Logistic Regression, SGD Classifier, RF Classifier, AdaBoost Classifier, and MLP, was first trained using a minimum set of features. Second, they suggested a deep learning architecture, contextual LSTM, to discriminate between tweets made by actual people and those generated by bots. The design of contextual LSTM incorporates both tweet text and account metadata. It is a system with various inputs and outputs that produces accurate categorization results.

Alom *et al.* [30] also proposed two deep learning techniques using Convolutional Neural Networks (CNNs) for identifying spam on Twitter at both the account and tweet levels. First, they developed a text-based classifier composed of an Embedding and a CNN layer to determine whether or not a particular tweet belongs to a spammer. Next, they suggested a combined classifier that utilizes both a text-based classifier and a neural network on users' information for identifying spammers at the account level on Twitter. For their tests, they used two Twitter datasets and compared the performance of their proposed machine learning and deep learning-based techniques to that of current state-of-the-art machine learning and deep learning-based approaches.

Wei and Nguyen [31] used a deep learning architecture consisting of an Embedding layer, three Bidirectional LSTM layers, and a fully linked layer to produce the final output for identifying whether tweets on Twitter were created by actual individuals or bots. They attained performance comparable to that of current cutting-edge bot detection systems.

Martinelli *et al.* [32] developed a simplified deep learning method for determining if a single tweet was produced by a spammer or not. In the tests, the authors developed many MLP classifiers with a range of zero to four hidden layers. As features (inputs to MLP classifiers), word embeddings were used. After loading pre-trained word embeddings, they specifically turned each word to a numerical vector and then averaged all words in sentences-tweets.

Gong *et al.* [33] proposed a more complex deep learning architecture and feature extraction approaches for detecting fraudulent users on Dianping, a location-based social network. First, they retrieved information that may be categorized into five major groups: time-series, spatial-temporal, user-generated content, social, and demographic aspects. The time-series characteristics were then used as input for the deep learning model, which consisted of a BiLSTM layer followed by a fully connected layer with a softmax activation function. This model's output consists of two probabilities (probability of legitimate and probability of malicious). The probabilities were then employed with the other data (traditional features) to train machine learning algorithms and get the final result. Multiple machine learning techniques, including XGBoost, RF, C4.5 Decision Tree, and SVM, were taught. According to the F1-score, XGBoost produced the best classification results.

In their research, Gilani *et al.* [34] classified Twitter accounts into two categories: Automated Bots and Real Users. They gathered data using their own platform, Stweeler. They gathered 2.5–3 million tweets every day and divided their data into four subsets: 10 million, one million, one hundred thousand, and one thousand, each representing the account's popularity based on the amount of followers. For the tagging procedure, they employed human annotation and Cohen's kappa coefficient to ensure that the annotator judgments were reliable. In all, 3536 accounts were applied in the testing phase throughout the four bands. The authors retrieved 15 characteristics and used the RF classifier after completing a statistical computation. They did 5-fold cross-validation by teaching and testing in three different sets of experiments. The accuracy rate was 86.44%, the precision was 85.44%, the recall was 82.24%, and the F-measure was 83.4%. Among the 15 traits, they discovered that six rated the highest. There are two issues with this study. First, it relies on humans. Second, it did not use the content as one of the attributes while using NLP for content analyzing may enhance the accuracy level of the system.

In a further recent work, Pozzana and Ferrara [35] examined four tweet metrics to determine how bots behaved during a single activity session: the number of mentions per tweet, the distance of the text in the tweet, the percentage of retweets, and the portion of answers. This study identified behavioral distinctions between human users and bot accounts that may be utilized to enhance bot detection algorithms. For example, humans are continually visible to tweets and messages from other users when engaged in online activities, boosting their chance of engaging in social contact. The authors employed five machine learning methods (Extra Trees (ET), DT, RF, Adaptive Boosting (AB), and KNN) to assess whether tweets were created by a bot or a human. The studies used a dataset

of over 16 million tweets posted by over 2 million unique individuals. ET and RF had the greatest cross-validated average performance 86% followed by DT and AB 83% and KNN 81%. However, the research's failure to categorize whether the bot is harmful or not might be seen as a flaw.

To detect spam-bots, Wang [36] employed three graph-based and three tweet-based features. The graph-based elements (such as the user's number of friends, followers, and follower ratio) are retrieved from the user's social network, whereas the tweet-based elements (such as the number of duplicate tweets, HTTP links, and replies/mentions) are retrieved from the user's most recent 20 tweets. The dataset applied to evaluate this approach includes 25,847 persons, around 500K tweets, and approximately, 49M followers/friends taken from publicly accessible Twitter data. Several classification techniques, including Decision Tree (DT), Neural Network (NN), Support Vector Machines (SVM), Naive Bayesian (NB), and k-Nearest Neighbors, are used to detect spam bots (KNN). With 91% accuracy, 91% recall, and 91% F-measure, the NB classifier achieved the best outcomes.

Chu *et al.* [37] classified Twitter users into three groups based on attributes retrieved from tweet content, tweeting behavior, and account proprieties: bot, human, and cyborg. The authors thought that bot character is less sophisticated than human behavior. They used an entropy rate to identify the difficulty of a process, with low rates indicating a regular process, medium rates indicating a difficult process, and high rates indicating a random process. The body of the tweet is utilized to create text patterns of recognized spam on Twitter. Other account-related factors, for example the percentage of external URLs, the safety of links, the date of account registration, and so on, are also applied in the classification. The RF machine-learning algorithm is applied to assess these factors to determine whether a Twitter account is a human, bot, or cyborg. The classifier's effectiveness is tested using a dataset of 500,000 different Twitter users. The total true positive rate for this strategy was 96.0% on average.

In addition, following multiple experiments, Ferrara *et al.* [38] generated an artificial intelligence program to spot bots on Twitter depending on variations in patterns of tasks among legitimate and fake accounts. They examined two distinct data sets of Twitter users who were grouped as bots or humans manually and by a pre-existing methods. The manually validated data collection included 8.4 million tweets from 3500 human accounts and 3.4 million tweets from 5000 bots. According to the study, human users reacted to other tweets 4 to 5 times more often than bots. Over the course of an hour, genuine users become more engaged, with the proportion of responses growing. The length of human users' tweets decreased as the sessions went. According to Ferrara, the quantity of information conveyed is decreasing. The author privileges that the change is related to cognitive tiredness, which causes individuals to be less inclined to exert mental effort in developing new material over time. Bots, in contrast, exhibit no change in their engagement or the quantity of material they tweet time to time.

5. CONCLUSION

This research covers papers on the analysis of real-time Twitter data, including classification and identification of bots and real-time sentiment analysis. To do this, the literature on Twitter sentiment analysis and bot identification and classification was analyzed. In addition, the research evaluates Twitter's platform characteristics, Streaming API, and data analysis stages.

According to the publications examined for this study's sentiment analysis, several academics have used opinion analysis to determine the negative and positive feelings of Twitter users. According to the studied articles, readers' sarcasm and irony were never effectively evaluated. According to the publications examined in this article, the length of tweets and a decrease in the amount of information communicated, which may be evaluated by detecting the tweet's interactivity, are patterns of behaviors that can be used to distinguish between actual and fraudulent Twitter accounts that this paper offers researchers with information on the categories of Twitter bots. In addition, the paper analyzes current Twitter analytic techniques and latest Twitter bot detecting systems. As a follow-up to this study, our feature research will use Twitter sentiment analysis to enhance bot detection classification.

REFERENCES

[1] D. M. Kancherla. "A Hybrid Approach for Detecting Automated Spammers in Twitter". *International Educational Applied Research Journal*, vol. 3, no. 9, 2707-2719, 2019.

[2] J. Rodríguez-Ruiz, J. I. Mata-Sánchez, R. Monroy, O. Loyola-González and A. López-Cuevas. "A one-class classification approach for bot detection on Twitter". *Computers and Security*, vol. 91, p. 101715, 2020.

[3] O. Loyola-González, R. Monroy, J. Rodríguez, A. L. Cuevas and J. I. Sánchez. "Contrast pattern-based classification for bot detection on twitter". *IEEE Access*, vol. 7, pp. 45800-45817, 2019.

[4] N. A. Azeez, O. Atiku, S. Misra, A. Adewumi, R. Ahuja and R. Damaševičius. "Detection of malicious URLs on twitter". *Advances*

in Electrical and Computer Technologies, vol. 672, pp. 309-318, 2020.

[5] J. Chen. "*Twitter Metrics: How and Why You Should Track Them"*. Sprout Social, United States. 2021. Available from: https://sproutsocial.com/insights/twitter-metrics/. [Last accessed on Nov 2021 23].

[6] S. Arifuzzaman and N. S. Sattar. "COVID-19 vaccination awareness and aftermath: Public sentiment analysis on twitter data and vaccinated population prediction in the USA". *Applied Sciences,* vol. 11, no. 14, p. 6128, 2021.

[7] O. Inya. "Egungun be careful, na Express you dey go: Socialising a newcomer-celebrity and co-constructing relational connection on Twitter Nigeria". *Journal of Pragmatics,* vol. 184, pp. 140-151, 2021.

[8] H. Piedrahita-Valdés, D. Piedrahita-Castillo, J. Bermejo-Higuera, P. Guillem-Saiz, J. R. Bermejo-Higuera, J. Guillem-Saiz, J. A. Sicilia-Montalvo and F. Machío-Regidor. "Vaccine hesitancy on social media: Sentiment analysis from June 2011 to April 2019". *Vaccines,* vol. 9, no. 1, p. 28, 2019.

[9] A. C. Breu. "From tweetstorm to tweetorials: Threaded tweets as a tool for medical education and knowledge dissemination". *Seminars in Nephrology,* vol. 40, no. 3, pp. 273-278, 2020.

[10] C. Wukich. "Connecting mayors: The content and formation of twitter information networks". *Urban Affairs Review*, vol. 58, pp. 33-67, 2020.

[11] N. Aguilar-Gallegos, L. E. Romero-García, E. G. Martínez-González, E. I. García-Sánchez and J. Aguilar-Ávilaa. "Dataset on dynamics of coronavirus on twitter". *Data in Brief*, vol. 30, p. 105684, 2020.

[12] S. Boon-Itt and Y. Skunkan. "Public perception of the COVID-19 pandemic on twitter: Sentiment analysis and topic modeling study". *JMIR Public Health Surveill,* vol. 6, no. 4, p. e21978, 2020.

[13] V. Cheplygina, F. Hermans, C. Albers, N. Bielczyk and I. Smeets. "Ten simple rules for getting started on Twitter as a scientist". *PLoS Computational Biology,* vol. 16, no. 2, p. e1007513, 2020.

[14] R. Chandrasekaran, V. Mehta, T. Valkunde and E. Moustakas. "Topics, trends, and sentiments of tweets about the COVID-19 pandemic: Temporal infoveillance study". *Journal of Medical Internet Research,* vol. 22, no. 10, p. e22624, 2020.

[15] P. Surowiec and C. Miles. "The populist style and public diplomacy: Kayfabe as performative agonism in Trump's Twitter posts". *Public Relations Inquiry,* vol. 10, no. 1, pp. 5-30, 2021.

[16] I. A. Mohammed and A. S. Abbas. "Twitter APIs for collecting data of influenza viruses, a systematic review". *2021 International Conference on Communication and Information Technology (ICICT)*, vol. 12, pp. 256-261, 2021.

[17] S. Wu, M. A. Rizoiu and L. Xie. "Variation across scales: Measurement fidelity under twitter data sampling". *Fourteenth International AAAI Conference on Web and Social Media*, vol. 14, no. 1, pp. 715-725, 2020.

[18] H. Ledford. "How Facebook, Twitter and other data troves are revolutionizing social science". *Nature*, vol. 582, no. 7812, pp. 328-330, 2020.

[19] Z. Pehlivan, J. Thièvre and T. Drugeon. "Archiving social media: The case of Twitter". *The Past Web*. Springer, Cham. pp. 43-56, 2021.

[20] I. Nazeer, S. K. Gupta, M. Rashid and A. Kumar. "Use of novel ensemble machine learning approach for social media sentiment analysis". *Analyzing Global Social Media Consumption*. Information Science Reference, Hershey. pp. 61-28, 2020.

[21] R. Al Bashaireh, M. Zohdy and V. Sabeeh. "Twitter Data Collection and Extraction: A Method and A New Dataset, the UTD-MI". *ICISDM 2020: Proceedings of the 2020 the 4th International Conference on Information System and Data Mining*, pp. 71-76, 2020.

[22] R. P. Mehta, M. A. Sanghvi, D. K. Shah and A. Singh. "Sentiment Analysis of Tweets Using Supervised Learning Algorithms". *First International Conference on Sustainable Technologies for Computational Intelligence*. vol. 1045, pp. 323-338, 2019.

[23] A. S. Neogi, K. A. Garga, R. K. Mishraa, Y. K. Dwivedib. "Sentiment analysis and classification of Indian farmers' protest using twitter data". *International Journal of Information Management Data Insights*, vol. 1, no. 1, p. 100019, 2022.

[24] T. Swathi, N. Kasiviswanath and A. A. Rao. "An optimal deep learning-based LSTM for stock price prediction using twitter sentiment analysis". *Applied Intelligence*, vol. 52, pp. 13675-13688, 2022.

[25] A. D. Dubey. "Twitter sentiment analysis during COVID-19 outbreak". *Jaipuria Institute of Management*, vol. 9, pp. 71-76, 2020.

[26] C. Monica and N. Nagaraju. Detection of fake tweets using sentiment analysis". *SN Computer Science,* vol. 1, no. 2, p. 89, 2020.

[27] G. A. Ruz, P. A. Henríquez and A. Mascareño. "Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers". *Future Generation Computer Systems,* vol. 106, pp. 92-104, 2020.

[28] S. E. Saad and J. Yang. "Twitter sentiment analysis based on ordinal regression". *IEEE Access,* vol. 7, pp. 163677-163685, 2019.

[29] S. Kudugunta and E. Ferrara. "Deep neural networks for bot detection". *Information Sciences*, vol. 467, pp. 312-322, 2018.

[30] Z. Alom, B. Carminati, E. Ferrarib. "A deep learning model for Twitter spam detection". *Online Social Networks and Media,* vol. 18, p. 100079, 2020.

[31] F. Wei and U. T. Nguyen. "Twitter Bot Detection Using Bidirectional Long Short-Term Memory Neural Networks and Word Embeddings". *2019 First IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA),* pp. 101-109, 2019.

[32] F. Martinelli, F. Mercaldo and A. Santone. "Social Network Polluting Contents Detection through Deep Learning Techniques". *2019 International Joint Conference on Neural Networks (IJCNN),* pp. 1-10, 2019.

[33] Q. Gong, Y. Chen, X. He, Z. Zhuang, T. Wang, H. Huang, X. Wang and X. Fu. "DeepScan: Exploiting deep learning for malicious account detection in location-based social networks". *IEEE Communications Magazine,* vol. 56, no. 11, pp. 21-27, 2018.

[34] Z. Gilani, E. Kochmar and J. Crowcroft. "Classification of twitter accounts into automated agents and human users". *Association for Computing Machinery*, vol. 17, p. 489-496, 2017.

[35] I. Pozzana, E. Ferrara. "Measuring bot and human behavioral dynamics". *Human Computer Interaction,* vol. 1, 1-11, 2018.

[36] A. H. Wang. "Detecting spam bots in online social networking sites: A machine learning approach". In: S. Foresti and S. Jajodia (Eds.), *Data and Applications Security and Privacy XXIV*. vol. 6166, pp. 335-342, 2010.

[37] Z. Chu, S. Gianvecchio, S. Jajodia H. Wang. Detecting automation of twitter accounts: Are you a human, bot, or cyborg?". *IEEE Transactions on Dependable and Secure Computing*, vol. 9, pp. 811-824, 2012.

[38] I. Pozzana and E. Ferrara. "Measuring bot and human behavioral dynamics". *Frontiers in Physics,* vol. 8, no. 125, p. 32, 2020.