Characterization of Dataset Exploiting Sentiment Analysis over Twitter

Uzaifa Siddiqui, Simra Ayaz, Maryam Nadeem, Sara Javed, Shahab Saquib Sohail

Department of Computer Science and Engineering, Jamia Hamdard, New Delhi, India

Corresponding author: Simra Ayaz, Email: simraayaz@gmail.com

Over the past many years in the world of social media, Twitter has attained a position as one of the most popular social networking sites (SNS) and provided to millions of people an opinion sharing platform. Thus, twitter started giving rise to many trending topics now and then and has attracted researchers for research in various dimensions. Since twitter is an opinion/sentiment sharing platform, sentiment analysis research has gained a lot of attention in recent years. These researches help in understanding the way people think and respond to an event as well as the purpose for which SNS are established. In this paper, we have outlined the most significant and relevant research in the above field in the recent few years. A novel characterization of twitter research data is presented and dataset creation and its implications are diagrammatically illustrated. The insight of dataset discussion would lay a strong foundation for the research in the concerned field. Moreover, it is envisaged that the future direction highlighted in the manuscript shall open various new dimensions of social network research.

Keywords: Twitter, Dataset, API, Sentiment Analysis, SNS.

2021. In Rahul Srivastava & Aditya Kumar Singh Pundir (eds.), *New Frontiers in Communication and Intelligent Systems*, 215–226. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.52458/978-81-95502-00-4-24

1 Introduction

With the expanding usage of social media add up due to the internet boom, there has been an upsurge in the utilization of micro blogging platform '*Twitter*'. The twitter users, also known as '*twitteratis*' express their perceptual opinions and their sentiments on and about vast range of topics such as elections [3], [10], [17], [45], [31], social and public issues such as demonetization [33], brands such as Starbucks [39] and to pretty much everything in 140-character broadcast / one to one message called 'tweets'. The openness and availability of these 140-character messages have been contributing to building up a rich dataset for research that has led to production of an enormous number of academic papers. One of the most common of these is '*Twitter sentiment analysis*', largely because of academic reasons and also because of brands and organizations wanting to know the review of their audience on their products and services, due to the capability of these broadcast messages in explaining and predicting business phenomena.

The fair share of research in the field of Twitter Sentiment Analysis has built up a huge dataset. We analyzed 50 recent sentiment analysis research studies conducted from 2011 to 2021 (till March) all aimed to review or implement different sentiment analysis solutions. In this paper, our contribution can be summarized as follows:

- State of the art review of the Twitter dataset with a special focus on sentiment analysis.
- We have constructed a filtered database of the relevant research in the field of sentiment analysis.
- A critical analysis of the dataset which provides a) extraction categories of the twitter data and b) characterization of these datasets with respect to the methodology and scope.
- Future direction for the research communities to take further the sentiment analysis research over social network platforms especially Twitter.

Rest of the paper is organized as follows: Section 2, Background focuses on the researches that are made around the similar topic and our contribution to it. Section 3, Methodology for acquiring dataset, focuses on the extraction and build-up of our database used in this paper. Section 4, Dataset Description and Characterization focuses on analysis of querying of multiple datasets through API and their characterization and reusability. And finally, we conclude in Section 5 with some possible future work.

2 Background

Since the launch of the micro blogging platform ***Twitter*** in 2006, its user base has been increasing ever since, and moreover after the internet boom era, in some countries twitter has undergone particularly extensive user adoption and fast growth in communication volume. The users communicate through micro broadcast messages called ***tweets***. Tweets provide much valuable insights, that if analyzed can be of great value, therefore the tweets database has been a great area of interest for the researchers, as it is also relatively easy to process, store and is accessible by its own 'Twitter API' and a few more third-party websites.

Twitter has also been the subject of many recent researches on Sentiment analysis, as tweets often express the user's opinion on a topic of interest, and therefore a huge and diverse amount of knowledge can be extracted from the tweets such as political opinions of people [23], views on evergoing debates on comparison between two products such as iphone vs samsung [27], organizations comparison on preference, such as of airlines [36] and reviews on products such as Google nest [20]. Twitter-based studies have constantly been advancing, and have experienced an escalation of research development in the last decade. To demonstrate this development a few review researches have been reported in the related field.

Williams et al. (2013) [52] first performed an analysis of 500 papers based on Twitter and related technology of micro blogging of papers between 2007 and 2011 on the basis of their abstract. The classification of filtered out papers was done across three dimensions-Aspect, Method and Domain. It concluded that many authors tend not to include the data corpus size and technology aspect in their abstract. Zimmer et al. (2014) [51] reviewed 380 scholarly articles, this was based on a full text analysis, inspired by Williams et al. (2013), and concluded about findings of the analysis methods, tweets and users under analysis. It also remarked on how fast the data was growing as they found 100 new articles in the first half alone of 2013.

These works lay out a set of knowledge bridges, entirely depicting the state of research on Twitter sentiments. By utilizing the knowledge provided in these papers, we build upon them our work, in which we focus more on the dataset utilized in the papers we collected, the data collection and data extraction techniques used in them. Williams et al. (2013), also considered a fourth aspect "data", which could not be included because of insufficient data in abstracts, so in this paper we perform a full analysis of Dataset used in twitter sentiment Analysis research papers. Our study focuses more on the dataset, the extraction API used by the researchers and the corpus, which makes our paper the first to do this study. Zimmer et al. (2014) concluded in their study about how fast the twitter dataset is growing and reforming and therefore we have used recent studies done, in order to inspect the revised data.

This study carries out a profound analysis of the dataset which lays out extraction techniques and characterization of the dataset, as regards to the scope of methodologies. In contrast to the previous articles, here our concern is not only the chronological growth of the twitter research, rather we focus on datasets, its diversity and possible inclusion of future research which shall eventually benefit the research community to take twitter research further.

3 Methodology of Acquiring Datasets

In this section, the methodology of collecting the dataset for this research study is discussed. The utilized dataset comprises ***50*** papers of the last 11 years, from 2011 to 2021 (till March) based on Twitter Sentiment Analysis generated by ***Twitter database***. The ***keyword*** used to access the papers from the database was ***'Twitter Sentiment Analysis'.***

From the numerous papers presented, more than 100 of those which were published from the year 2011 to 2021 (till March) were collected and those of which clearly exhibit precise and accurate dataset of tweets and focused mainly on Sentiment Analysis were filtered out. Our team first filtered the irrelevant papers (papers without 'sentiment analyses in their title or mentioned keyword) and those papers which do not have any significance to related research. Our team dedicated weeks to perform all these sorting. Finally, we came up with a total of 50 papers of the same which manifests Sentiment Analysis on tweets in different areas. These papers were then stored in our database to collect more information from them to conclude in our paper for more results.

The total no. of papers of each year is shown in the table below with the maximum number of papers on Sentiment Analysis in 2018. Figure.1 shows the steps taken to build our dataset. Table 1 represents the number of twitter sentiment analysis researches that took place in each year from 2011 to 2021 (till March) in our database. Further in Figure 2, a graph is also shown for the same. Although we could add a few more papers, to respect the page limit, we have restricted this manuscript to 50 papers, and we intend to take this research further for its extended version.

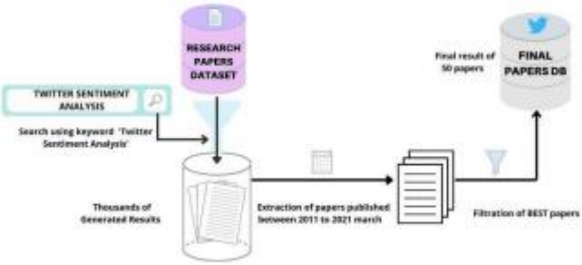
RESEARCH PAPERS DATASET
TWITTER SENTIMENT ANALYSIS
Search using keyword 'Twitter Sentiment Analysis'
Thousands of Generated Results
Extraction of papers published between 2011 to 2021 march
Filtration of BEST papers
Final result of 50 papers
FINAL PAPERS DB

Fig. 1. The construction of our database of Twitter Sentiment Analysis Research papers published between 2011 to 2021

Table 1. Total number of researches on twitter sentiment analysis in each year from 2011 to 2021 out of the selected 50 papers

Year	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Total no. of papers	02	03	02	05	04	05	01	11	07	06	04

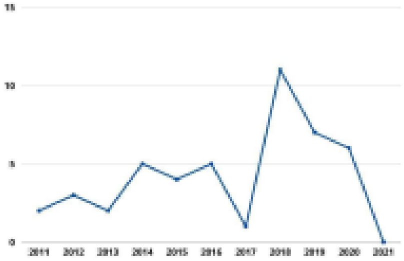

Fig. 2. Number of papers selected from each year between 2011 to 2021

4 Dataset Description and Characterization

Twitter sentiment analysis involves a series of procedures to be followed step by step in order to achieve output with promising accuracy. Out of which, collection of most suitable data for the research is the very first and foremost step of the process. In this paper we have reviewed multiple researches on twitter sentiment analysis and discussed the various retrieval methods of tweets in the form of data that is to be processed further. The unstructured data retrieved is then pre-processed (by tokenization, normalization, POS tagging and other techniques) and using classification methodologies (like Naive Bayes classifier, Random Forest classifier, SVMs, etc) the data is trained and classified. In accordance with our paper, Figure.3 depicts the numerous ways and data types engaged in the retrieval process that we have discussed in the sections below. It briefly illustrates how the collected data (the tweets) were adopted by employing few of the techniques that we have explored in this paper.

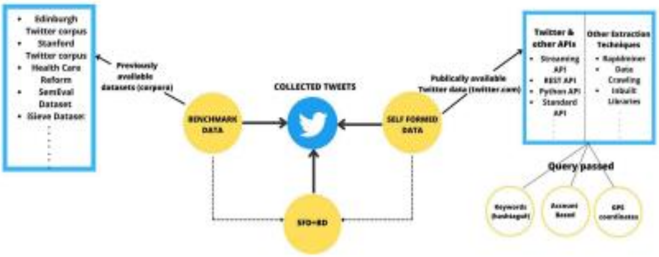

Fig. 3. Approach most commonly used for collecting data to analyze sentiments

4.1 Data collection techniques for sentiment analysis

A major step in any Twitter-based research project is data collection. In this section, we discuss techniques for data collection for analyzing sentiments using Twitter. We start by briefly describing Twitter's application programming interfaces (APIs) and by briefing its sub-APIs. Another API for gathering data other than tweets called Yahoo! Finance API has also been discussed.

We then discussed hashtag/keyword-based, location-based, account-based searches. These two approaches cover most data collection logics on Twitter.

4.1.1 Twitter API

The Twitter APIs makes it easy to collect a large number of tweets that have certain characteristics such as those posted by a particular user or from a particular location or those consisting of some specific terms or emoticons. Twitter offers two types of application programming interfaces (APIs): REST and Streaming. The REST-API allows developers access to read and write Twitter data. For researchers, this API is a valuable access point to hunt for data posted in the recent past. On the other side, the Streaming API allows developers to access Twitter's stream of data in real-time. This means that the user can obtain the latest tweets which contain specific terms or the tweets that were posted by some specific users. It has been found out that most researchers use Streaming API more due to the real time access to tweets.

4.1.2 Yahoo! Finance API

Yahoo Finance API is used to provide historical and real-time data about stocks, crypto currencies, bonds, and information on financial markets and products.

It is used by researchers for collecting data that they need to incorporate in their study. In most cases, the data prices of stocks that are extracted over a range of periods to create datasets for further investigations like sentiment analysis.

Our corpus included 50 studies that were also used in explicitly analyzing how the data was obtained. Table 2 illustrates various extraction techniques and APIs used by the researchers and its percentage distribution in accordance with our data. Researchers applied a variety of approaches to acquire data for advocating further research. It is clear from table 2 that Streaming API was individually the most used tool with 22% of the studies working with it. REST API is the second most utilized approach for collecting tweets. Other APIs including Twitter Search API as appropriated by Ptacek et al. [11], Twitter

Standard API used by Gabarron et al. [34] for gathering tweets. There are other techniques to collect tweets like APIs in certain languages like Python [43], R Language [42], and some researchers incorporated web scraping and crawling techniques to obtain data [35]. Apart from the extraction of tweets, other data like stock values are obtained by Yahoo! Finance API as done by Mittal et al. [2]. In all, 4 studies did not mention the extraction technique used to attain the datasets.

Table 2. Usage of various APIs and extraction techniques adopted by researchers to extract data

	Only REST API	Only Streaming API	Rest + Streaming	Yahoo API	Other APIs	Other Extraction Techniques	Not Mentioned
No. of papers	5	11	1	2	22	5	4
Percentage	10%	22%	2%	4%	44%	10%	4%

4.2 Search Query for extracting tweets

Twitter provides access to its large amount of data which is publicly available for the researchers to openly crawl. A simple approach for data collection on Twitter using APIs is by selecting tweets using appropriate character strings as keywords/hash tags. The query chosen solemnly focuses on extracting the most relevant sets of tweets for utter precision in analysis performed. The queries passed should be relevant to the study conducted by the researchers. Another way of assembling target data is by selecting tweets based on their authors. The databases are queried based on the accounts of users that are appropriate to the research, for example, celebrities, politicians, people belonging to a certain community, people with official accounts, etc. Location-based search queries are also another way to extract tweets of the target area around which the analysis revolves.

McIver et al. [14] analysed sleep issues using TSA (Twitter Sentiment Analysis) by collecting tweets with the presence of keywords like insomnia, "can't sleep", Ambien, and more and hashtags like "#teamnosleep", and "#cantsleep". Likewise, Hasan et al. [23], Menendez et al. [25], Dhanya et al. [33], Alomari et al. [37], Pinto et al. [38], Gracia et al. [47] and most of the researchers in our dataset drew their pertinent tweets by querying with a keyword or hashtags as per the requirement of the research. In contrast, few of the researchers instead targeted tweets with GPS coordinates, like, Chen et al. [15] generated the twitter posts by coordinating Twitter streaming API with coordinates [- 87.94, 41.64] (South-West limit) and [-87.52, 42.02] (NorthEast limit) within the boundaries of Chicago. Similarly, Lim et al. [29] collected tweets geo-tagged with latitude/longitude coordinates posted under the 5km × 5km grid of central Melbourne, Australia. With the best of our knowledge, we have also observed that Wang et al. [3] along with keywords as their query also extracted the subject matter of presidential election by retrieving tweets with candidate name tags including names with common typos like for Mitt Romney, @MittRomney, @PlanetRomney, @MittNews, @believeinromney were used.

4.3 Corpora and Twitter

With the assessment of APIs, a unique dataset can be established in the form of a corpus. Fundamentally, a corpus is a collection of large texts (written or spoken) which is stored in the form of data in a computer database. These corpora can be built of various sizes and for several motives, out of which, analysis performed on its data is an extensive approach. Here, in this paper we have discussed the corpus/corpora made for conducting sentiment analysis on twitter data.

For constructing these datasets, APIs like Twitter API are queried to return required particulars. They allow easy access to crawl a large number of tweets consisting of particular keywords, emoticons or hash tags. The derived data thus can be used as a corpus by the researchers to perform their analysis. In this paper, we have distinguished and identified many prominent datasets, commonly used by researchers as a source of extraction for their unique corpus or as a whole. With the increasing trend in TSA a number of evaluation datasets have been built. These evaluation datasets comprise annotated sets of tweets with tweet's sentiment. In this paper, we have observed from our dataset that 20% of researches conducted have used evaluation datasets, i.e., pre-existing datasets as a source of their analysis. In the section below we have briefly described some of the famously available evaluation datasets made for twitter sentiment analysis.

4.3.1 Edinburgh Twitter Corpus (ETC)

ETC [54] is one of the most recognized datasets for sentiment analysis consisting of about 96 million tweets made out of over 2 billion words. The data was collected by crawling through Twitter's Streaming API from the span of November 11th 2009 until February 1st 2010 turning out to be the representative sample of tweets for the entire period. A number of researchers till date are using this corpus for carrying out various strains of sentiment analysis like Kouloumpis et al. [1] created their HASH dataset by filtering out duplicate tweets, non-English tweets, and tweets without hashtags and specified 15 most-used hashtags in the Edinburgh corpus and created their unique dataset along with other corpus like iSieve dataset for evaluating the effectiveness of the features from section for sentiment analysis in Twitter data.

4.3.2 Stanford Twitter Sentiment (STS)

The STS corpus [55] serves with two types of sets: training sets and testing sets. Tweets containing at least one emoticon were collected from April 6 till June 25, 2009. The tweets were then automatically annotated as positive (containing emoticons like :), :-), :), :D, or =)) or as negative (containing :(, :-(, or : ( ). A sum of 1.6 million annotated tweets as training set was cumulated and 182 positive and 177 negative tweets were manually annotated for training set.

4.3.3 Health Care Reform (HCR)

Speriosu et al. [56] manually annotated the dataset HCR which comprises 2,515 tweets classified into 541 positive, 1,381 negative, 470 neutral, 79 irrelevant, and 44 unsure tweets. On the topic her in March 2010 the collected tweets were manually annotated for polarity aiming at one of the following: health care reform, Obama, Democrats, Republicans, Tea Party, conservatives, liberals, and Stupak and were distinguished into 3 sets the training, the development, and the testing, where each consisted of about 840 tweets.

In continuation with the subject, broaching a few more benchmark datasets that researchers have opted for in their papers. Kouloumpis et al. [1] along with Edinburgh corpus included the iSieve dataset which approximately contains 4,000 tweets. Venkit et al. [49] used SemEval dataset which was collected by geolocalzation and incorporated it in their own self formed data entering in the 8% of papers which covers both, as depicted in table 3. Pla et al. [9] pre-processed the General Corpus of TASS2013 containing almost 68000 tweets written in Spanish. These tweets were collected from the accounts of 150 well-known personalities and celebrities of the world of politics, economy, communication, mass media and culture, posted from November 2011 till March 2012.

4.4 Benchmark Data vs Self-formed Data

Datasets required to carry out analysis or for other purposes can be categorized as self-formed and benchmark datasets. Self-formed datasets are built by people involved in the task. Such datasets are formed entirely from scratch and fundamentally need to go through several stages like pre-processing to become utilizable. Whereas, benchmark datasets or public datasets are readily available and don't need to be any kind of refining or pre-processing from scratch.

In sentiment analysis using Twitter, there are a couple of publicly available corpora that are obtainable from Kaggle and other such resources. Some of these datasets are discussed in section 4.3 that can be used by researchers for their analysis. However, in some cases where the set of tweets in these are not relevant to the study, self-made datasets are favoured. These newly created datasets by people are also eligible to be uploaded to the internet to support reusability.

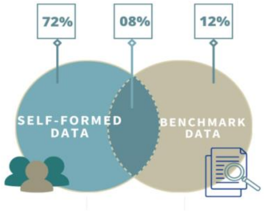

Fig. 4. Venn diagram depicting percentages of self-formed data, benchmark data and both (remaining 8% of not mentioned data)

In our analysis, it was found that most of the researchers had built their own datasets (shown in figure 4) as done by Sulthana et al. [31], Dhanya et al. [33], Schumaker et al. [21], and so on.

On the contrary, some studies had deployed benchmark datasets since there was no requirement for explicitly creating data for sentiment analysis as used by Ankit et al. [54].

Apart from these two categories, some research works included both benchmark as well as self-created datasets carried out by Bian et al. [20]. However, in some researches like [24], the sources of data that they had used for performing analysis were not mentioned.

Table 3 shows the percentage distribution of the type of datasets adopted by researchers in our study which clearly determines that self-formed datasets are most commonly considered as an option to perform research as compared to benchmark datasets.

Table 3. Percentage distribution of types of datasets found during the analysis

	Self-Formed Data (SFD)	Benchmark Data (BD)	SFD + BD	Not mentioned
Total	36	6	4	4
Percentage	72%	12%	8%	8%

4.5 Reusability of Datasets

While the analysis was being conducted, it was observed that a few researchers had mentioned in their studies that their dataset can be reused in the future. Sanders et al. [48] have mentioned that their corpus can be obtained from GitHub. Similarly, Gabarrom et al. [34] have made the tweets that were collected, to be publicly available by publishing on the Internet, whereas, this information was not at all mentioned in most of the papers in our corpus.

5 Conclusion and Future Work

The massive number of discrete types of data present on social media captivates the curiosity of researchers. Twitter has been one of such platforms to have been studied thoroughly. Various researches have been made that recognize the valuable cognizance that can be extracted from it, and a few researches have also been made that review this development. This study takes into consideration relevant and significant Twitter sentiment analysis papers, published between 2011 and 2021 (till March) that aimed to implement or review sentiment analysis solutions. The study is focussed on the datasets used in these research articles, their data collection techniques, the corpus used in them and the APIs used to extract the tweets, and finally we performed a comparative study based on different datasets and APIs used.

The study takes heed of the corpus used, in which it is remarked that in 8% of the papers we could not identify which of the dataset extraction technique was used and almost 12% of the researches used pre-defined corpus (bench-mark data), 72% of the researches were based on self-defined corpus (self-formed data) and 8% of the researches implicitly used both the Self-formed data and bench-marked data. Since majorly, the dataset is self-formed, therefore we took into consideration their extraction techniques, the APIs used to extract them.

In future, a detailed study can be performed analysing the dataset, the user space and the field of applications in which the researches have been done. We can take a bigger amount of twitter sentiment analysis papers published over the years, to review the insights. We can take into consideration more third-party APIs and extraction methods. Moreover, how the users are interacting with these social platforms, their classifications and most relevant application area shall be our focus in our extended work for this article.

References

[1] Kouloumpis, E., Wilson, T. and Moore, J. (2011). Twitter sentiment analysis: The good the bad and theomg!. In *Proceedings of the International AAAI Conference on Web and Social Media*. 5(1): 538-541.

[2] Mittal, A. and Goel, A. (2012). Stock prediction using twitter sentiment analysis. Standford University, CS229, 15.

[3] Zhang, M. (2012). *Proceedings of the ACL 2012 System Demonstrations.*

[4] Rao, T. and Srivastava, S. (2012). Analyzing stock market movements using twitter sentiment analysis. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 119-123.

[5] Wakade, S. et al. (2012). Text mining for sentiment analysis of Twitter data. In *Proceedings of the International Conference on Information and Knowledge Engineering (IKE). The Steering Committee of the World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).*

[6] Neethu, M. S. and R. Rajasree. (2013). Sentiment analysis in twitter using machine learning techniques. In

Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT).

[7]Myslín, M. et al. (2013). Using twitter to examine smoking behavior and perceptions of emerging tobacco products. *Journal of Medical Internet Research*, 15(8): e174.

[8]Abeywardena, I. S. (2014). Public opinion on OER and MOOC: A sentiment analysis of twitter data. In the *International Conference on Open and Flexible Education.*

[9] Ferran, P. and Hurtado, L. F. (2014). Political tendency identification in twitter using sentiment analysis techniques. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical Papers.*

[10]Rill, S. et al. (2014). Politwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis. *Knowledge-Based Systems*, 69: 24-33.

[11] Haripriya, V. and Patil, P. G. (2017). A survey of sarcasm detection in social media. *International Journal for Research in Applied Science & Engineering Technology*, 5(12): 1748-1753.

[12]Widener, M. J. and Wenwen, L. (2014). Using geolocated Twitter data to monitor the prevalence of healthy and unhealthyfood references across the US. *Applied Geography*, 54: 189-197.

[13]Holmberg, K. and Hellsten, I. (2015).Gender differences in the climate change communication on Twitter. *Internet Research,* 5: 5-11.

[14]McIver, D. J. et al. (2015). Characterizing sleep issues using Twitter. *Journal of Medical Internet Research*, 17(6): e140.

[15]Chen, X., Cho, Y. and Jang, S. Y. (2015). Crime prediction using Twitter sentiment and weather. In *Systems and Information Engineering Design Symposium.*

[16]Vicente, M. et al. (2015). Twitter gender classification using user unstructured information. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE).*

[17]Ramteke, J. et al. (2016). Election result prediction using Twitter sentiment analysis. In *the International Conference on Inventive Computation Technologies (ICICT)*, 1.

[18]Alrubaian, M. et al. (2016). A credibility analysis system for assessing information on twitter. *IEEE Transactions onDependable and Secure Computing*, 15(4): 661-674.

[19]Sasank, P. V. et al. (2016). Sentiment analysis of Twitter data for predicting stock market movements. In *International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES).*

[20] Jiang, B. et al. (2016). Mining Twitter to assess the public perception of the “Internet of Things”." *PloS one*, 11(7): e0158450.

[21]Schumaker, R. P., Jarmoszko, A. T. and Labedz, C. S. (2016). Predicting wins and spread in the Premier League using a sentiment analysis of twitter. *Decision Support Systems*, 88: 76-84.

[22]Xia, L. et al. (2017). An investigation of brand-related user-generated content on Twitter. *Journal of Advertising*, 46(2): 236-247.

[23] Ali, H. et al. (2018). Machine learning-based sentiment analysis for twitter accounts. *Mathematical and Computational Applications*, 23(1): 11.

[24] Marco, F. and Montangero, M. (2018). Sentiment analysis and twitter: a game proposal. *Personal and Ubiquitous Computing*, 22(4): 771-785.

[25]Reyes-Menendez, A. et al. (2018). Understanding# World Environment Day user opinions in Twitter: A topic-based sentiment analysis approach. *International journal of environmental research and public health*, 15(11): 2537.

[26] Lee, C. et al. (2018). Investigating the emotional responses of individuals to urban green space using twitter data: A critical comparison of three different methods of sentiment analysis. *Urban Planning*, 3(1): 21-33.

[27] Hasan, M. R. Sentiment Analysis with NLP on Twitter Data. In *International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2)*, 1-4.

[28]Wagh, R. and Punde, P. (2018). Survey on sentiment analysis using twitter dataset. In *Second InternationalConference on Electronics, Communication and Aerospace Technology (ICECA).*

[29] Hui, L. K. et al. (2018). The grass is greener on the other side: Understanding the effects of green spaces on Twitter usersentiments. *Companion Proceedings of the Web Conference.*

[30] Das, S. et al. (2019). Extracting patterns from Twitter to promote biking. *IATSS Research*, 43(1): 51-59.

[31]Sulthana, A. R. et al. (2018). Sentiment analysis in twitter data using data analytic techniques for predictive modelling. *Journal of Physics: Conference Series,* 1000(1). IOP Publishing.

[32] Nabizath, S. (2018). An ensemble classification system for twitter sentiment analysis. *Procedia Computer Science*, 132: 937-946.

[33] Dhanya, N. M. and U. C. Harish. (2018). Sentiment analysis of twitter data on demonetization using machine learning techniques. *Computational vision and bio inspired computing*, 227-237.

[34]Gabarron, E. et al. (2019). Diabetes on Twitter: a sentiment analysis. *Journal of diabetes science and technology*, 13(3): 439-444.

[35]Debby, A., Priyanta, S. and Rokhman, N. (2019). Analysis of Emoticon and Sarcasm Effect on Sentiment Analysis of Indonesian Language on Twitter. *Journal of Information Systems Engineering and Business Intelligence*, 5(2): 100-109.

[36] Prabhakar, E., et al. (2019). Sentiment analysis of US airline twitter data using new adaboost approach. *International Journal ofEngineering Research & Technology*, 7(1): 1-6.

[37]Ebtesam, A., Mehmood, R. and Katib, I. (2019). Road traffic event detection using twitter data, machine learning, and apache spark. In *IEEE Smart World, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI).*

[38]Pinto, J. P. and Murari, V. (2019). Real time sentiment analysis of political twitter data using machine learningapproach. *International Journal ofEngineering Research & Technology,* 6(4):4124-4129.

[39]Shirdastian, H., Laroche, M. and Richard, M. O. (2019).Using big data analytics to study brand authenticity sentiments: The case of Starbucks on Twitter. *International Journal of Information Management*, 48: 291-307.

[40] Verma, P. et al. (2019). Twitter sentiment analysis on Indian government project using R. *Int J Recent Technol Eng,* 8(3): 8338-8341.

[41]Wathanti, S., Wirapong, C., and Tuamsuk, K. (2020). THAI CUSTOM INFORMATION SHARING ON THE INTERNET BY LINKED DATA TECHNIQUES. *Journal of Critical Reviews*, 7(8): 1398-1402.

[42]Sreeja, I., Joel V. S. and Jatian, L. (2020). Twitter Sentiment Analysis on Airline Tweets in India Using R Language. *Journal of Physics: Conference Series*. 1427(1). IOP Publishing.

[43] Manguri, K. H. et al. (2020). Twitter sentiment analysis on worldwide COVID-19 outbreaks. *Kurdistan Journal of Applied Research*, 54-65.

[44]Tzacheva, A. A., Ranganathan, J. and Bagavathi, A. K. (2020). Action rules for sentiment analysis usingTwitter. *International Journal of Social Network Mining*, 3(1): 35-51.

[45]Sharma, A. and Ghose, U. (2020). Sentimental analysis of twitter data with respect to general

elections in india. *Procedia Computer Science*, 173: 325-334.

[46] Albaldawi, W. S. and Almuttairi, R. M. (2020). Near Real Time Twitter Sentiment Analysis and Visualization. *IOPConference Series: Materials Science and Engineering,* 928(3). IOP Publishing.

[47]Klaifer, G. and Berton, L. (2021). Topic detection and sentiment analysis in Twitter content related to COVID-19 fromBrazil and the USA. *Applied Soft Computing*, 101: 107057.

[48] Sanders, A. C. et al. (2021). Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of COVID-19 Twitter discourse. *medRxiv*, 2020-08.

[49]Venkit, P. et al. (2021). ASourceful'Twist: Emoji Prediction Based on Sentiment, Hashtags and Application Source. *arXiv preprint arXiv*:2103.07833.

[50] Badgaiyya, A. et al. (2021). An Application of Sentiment Analysis Based on Hybrid Database of Movie Ratings.

[51] Bruns, A. et al. (2014). A topology of Twitter research: disciplines, methods, and ethics. *Aslib Journal of InformationManagement*.

[52] Williams, S. A., Melissa, M. T. and Warwick, C. (2013). What do people study when they study Twitter? Classifying Twitter related academic papers. *Journal of documentation*.

[53] Giachanou, A. and Crestani, F. (2016). Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys (CSUR)*, 49(2): 1-41.

[54] Petrović, S., Miles O. and Lavrenko, V. (2010). The edinburgh twitter corpus. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*.

[55]Go, A., Bhayani, R. and Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12).

[56] Speriosu, M. et al. (2011). Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the First workshop on Unsupervised Learning in NLP*.

[57]Zimbra, D. et al. (2018). The state-of-the-art in Twitter sentiment analysis: A review and benchmark evaluation. *ACM Transactions on Management Information Systems (TMIS)*, 9(2): 1-29.

[58]Fahad, A. N. and AlMansour, A. A. (2019). State-of-the-art review on Twitter Sentiment Analysis. In *2nd International Conference on Computer Applications & Information Security (ICCAIS).*

[59] Khan, M. et al. (2021). How They Tweet? An Insightful Analysis of Twitter Handles of Saudi Arabia. *arXiv e-prints (2021): arXiv-2105.*

[60] Saquib, S. S., Khan, M. M. and Alam, M. A. (2021). An Analysis of Twitter Users from The Perspective of Their Behavior, Language, Region and Development Indices--A Study of 80 Million Tweets. *arXiv e- prints (2021): arXiv-2105.*

[61] Saquib, S. S. et al. (2021). Crawling Twitter data through API: A technical/legal perspective. *arXiv preprint arXiv:2105.10724* .