JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A REVIEW ON SENTIMENT ANALYSIS APPLICATIONS AND APPROACHES

[1]D.SAI TVARITHA, [2]NITHYA SHREE J, [3]SAAKSHI NS
[4]SURYA PRAKASH S, [5]SIYONA RATHEESH, [6]SHIMIL SHIJO
[1]Student department of BCA,[2]Student department of BCA,[3]Student department of BCA,
[4]Student department of BCA,[5]Student department of BCA,[6]Assistant Professor department of BCA

[1]Jain[Deemed-to-be] University, Bangalore, India,[2]Jain[Deemed-to-be] University, Bangalore,India ,[3]Jain[Deemed-to-be] University, Bangalore, India,[4]Jain[Deemed-to-be] University, Bangalore, India,
[5]Jain[Deemed-to-be] University, Bangalore, India,[6]Jain[Deemed-to-be] University, Bangalore, India

Abstract: This generation completely depends on the text opinions and ideologies of the viewers for which they invest a lot of time on various social media platforms such as Instagram, Twitter, and Facebook to share their thoughts and get opinions on the same. This method of sharing their knowledge and emotions with society and social media drives businesses to gather more information about their companies, products, feedback and how well known they are among the people allowing them to make more important business decisions. Social media is a platform in which an abundance of data is generated from various resources which can partially be understood or not. This has made way for technology that helps in language processing of data making it user friendly. Here we emphasize text representation, as emotions that play a vital role as a response to a shared post. Social media data helps individuals and businesses to take decisions based on rigorous data analysis. Massive amounts of data are generated by users in the forms of opinions, reviews, emotions, arguments, viewpoints, and so on about various social events, products, brands and politics, movies, and so on. The importance of Twitter sentiment analysis in discovering similar text patterns in the given input text cannot be overstated. Furthermore, this classification results in positive, negative, and neutral evaluations. Also, different approaches to perform sentiment analysis like Machine Learning, Lexicon-based, Naive Bayes algorithm and deep learning techniques are discussed here.

***IndexTerms* -** Natural language processing, Naive Bayes Algorithm, Machine Learning, Deep Learning.

I. INTRODUCTION

Sentiment analysis is a part of Natural Language Processing (NLP) that includes the study of the intensity of the reactions conveyed in a form of textual data. The automated analysis of the messages delivered through social media is one of the research fields, both in academics and in industry because of its wide usefulness in various domains[17]. Sentiment Analysis is a popular application in text mining, and with the integration of machine learning algorithms and deep learning algorithms, it becomes more effective and is used in a wide range of businesses to increase productivity and provide a better customer experience. Sentiment analysis deals with text and it is based on Subjective context that mainly focuses on text analysis, natural language processing, computational linguistics and biometrics to systematically identify, extract, quantify and study subjective information. The research on sentiment analysis methodologies are progressing everyday due to the easy availability of huge amount of raw data which is generated by social medias, blogs, forums etc. Social Media platforms like Facebook, Twitter, Instagram are generating millions of status updates, posts and tweet messages in every minute which reflects people's opinions and attitude towards particular topic[17].

In this paper, the users learn people's opinions, attitudes, and emotions towards the given posts. Lexicon based language is characterized as a polarity decision to check the words that can help the system to categorize as positive, negative and neutral tweets. Sentiment analysis is used by various parties for the marketing of products, used by public figures to analyze their activities in order to gain followers and views respectively. On Twitter, people can express their views by posting tweets on a variety of topics. This application can be used to analyze these posts and then determine whether they are viewed positively or negatively by the audience[7]. In order to understand the importance of public sentiments and the market value of companies, there is a need for an analytics tool, wherein businesses can estimate the direction of marketing by analyzing the polarity of the received comments. For this purpose, data gathering has to be done using Twitter APIs. It will be followed by data preprocessing which involves the removal of stop words, empty fields, URLs, hashtags etc. and stemming. Later, the preprocessed data has to be trained using a sentiment classification model(model building) and finally accuracy of the model is to be tested using test data. After testing, data will be checked and polarity will be assigned[20].

II. PHASES OF SENTIMENT ANALYSIS

To perform Sentiment analysis, data gathering has to be done initially from the source(Twitter) which is followed by various stages of preprocessing. Then sentiment analysis is performed with the help of suitable techniques and datasets. Each data set is divided into three categories to get the desired output. They are, 1. Training Data Sets: it is used into progressing machine learning in the algorithm on the learning process. 2. Testing Data Sets: it is used to see if our algorithm is overfitting or not. 3. Validation Data Sets: it is used to assess the execution of the algorithm.

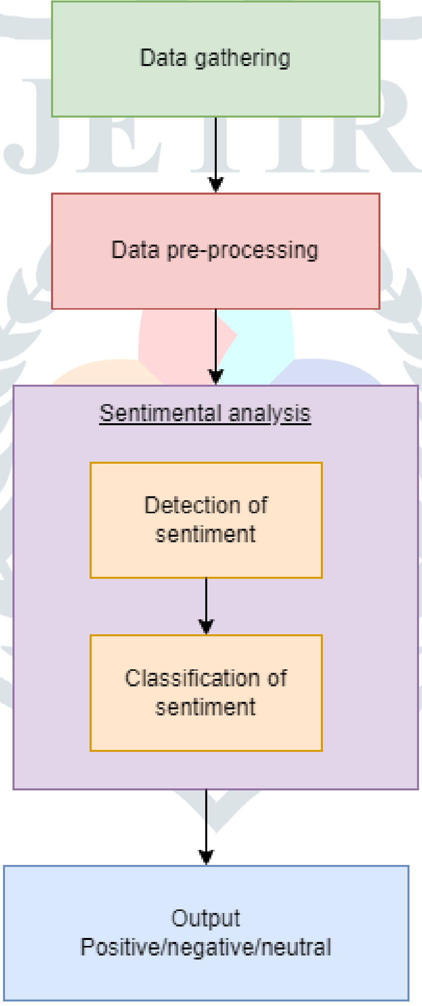


figure 1 phases of sentiment analysis

A. Data Gathering: The first step is to collect quality. In case of Twitter, Tweet collection includes the collection of appropriate tweets on a chosen topic. These tweets are fetched with the help of Twitter APIs which act as an interface between the user and Twitter.

B. Data Pre-processing: The data collection stage only fetches the raw data with which sentiment analysis cannot be performed accurately. It is important the clean the data by removing irrelevant tokens and converting inti the root word. This phase is known as data preprocessing. This stage identifies the potential of the synthetical correlation among the tweets. Removal of unwanted tokens like usernames, hashtags, unwanted spaces, irrelevant special characters, stop words, abbreviation’s , URL’s etc. are also done in data pre-processing phase.

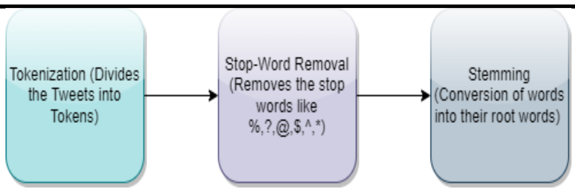

Tokenization (Divides the Tweets into Tokens)
Stop-Word Removal (Removes the stop words like %,?,@,$,^,*)
Stemming (Conversion of words into their root words)

figure 2 data pre-processing steps

C. Sentiment Analysis: This phase involves detection and classification of sentiment in pre-processed text. At this stage objective expressions can be eliminated and only subjective expressions has to be considered. The phase of sentiment analysis can be done through different computational techniques, model building and algorithms[6]. Later, these texts are categorised into positive, negative or neutral comments which is known as sentiment classification.

D. Output Last stage of sentiment analysis is the presentation of obtained output. Expected outputs is the polarity of the text. That can be positive, negative or neutral. Also output can be presented as visualization like graphs and charts.

III. APPROACHES OF SENTIMENT ANALYSIS

In this paper we have discussed four different approaches to perform sentiment analysis. They are (i). Machine Learning, (ii). Naive Bayes, (iii). Deep learning and (iv) Pre-trained and Rule based VADER models.

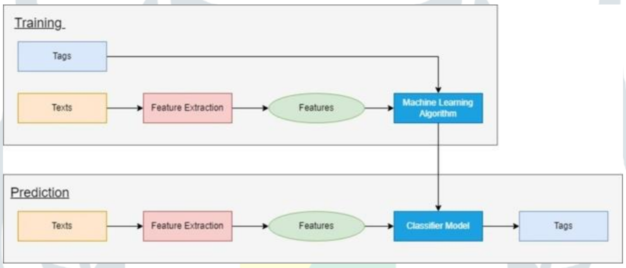


figure 3 Sentiment Approaches

i. Machine Learning for NLP

ML for NLP is in need of preprocessing to retrieve featuresfrom original text. Before the implementation of Machine Learning Models, the conversion of textual data into numerical representation is essential. It can be done through the processes like Vectorization. Supervised ML requires data to belabelled which means if there is any subjectivity in the content then that has to be returned to the model. In comparison with pre-trained models, custom models provide more manageability overtheoutcomeandare more apt for certain applications which are specific in nature.

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

equation 1Bayes theorem

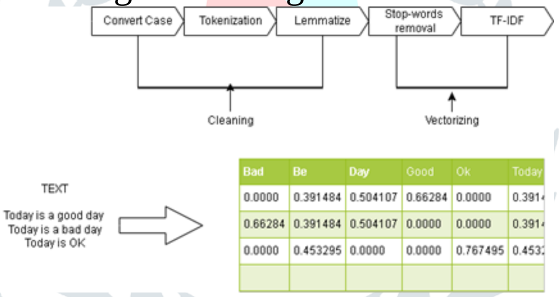
Working of bayes theorem in ML:

In machine learning, navies bayes classifiers are a family of simple 'probabilistic classifiers' based on applying Bayes theorem with strong (naïve) independence assumptions between the features. Using Bayes probability terminology, the above equation can be written as Posterior=prior*likelihood/evidence .

Naïve Bayes classifier is the simplest problematic model that includes positivity on text classification. Here , we worked with Bayes probability rule which have self supporting feature, that includes text classification which can be used for analysis of text data. We calculate the probability of each tag for a given text and then the tag with highest one.

ii. Naive Bayes

The Naive Bayes classifier is an algorithm that utilizes the chances of making predictions which are based on prior known knowledge of conditions which are associated each other. Along with that it also uses conditional probabilities of the lexical characters that is present in the positive or negative content of the training data. A simple DTM(Document Term Matrix) is designed initially for Naive Bayes. Construction of model requires additional information's/features like text length, named entities, time or publication location etc. DTM has a result in a white feature space as each distinct word or phrase in the lexicon that will be mapped to a numerical model using vectorizing model which will be identified by the system. Data preprocessing stage also include to perform reduction of dimensionality and classifiers. Also data pre-processing for NLP is seen in the vector representation which is constructed by counting the TF (Term Frequency) and waiting for the same with IDF (Inverse Document Frequency). To capture some context in text as transformation has to be applied for the generation of training and testing data.



Bad	Be	Day	Good	Ok	Today
0.0000	0.391484	0.504107	0.66284	0.0000	0.391
0.66284	0.391484	0.504107	0.0000	0.0000	0.391
0.0000	0.453295	0.0000	0.0000	0.767495	0.453

figure 4naive bayes data pre-processing weighted frequency count

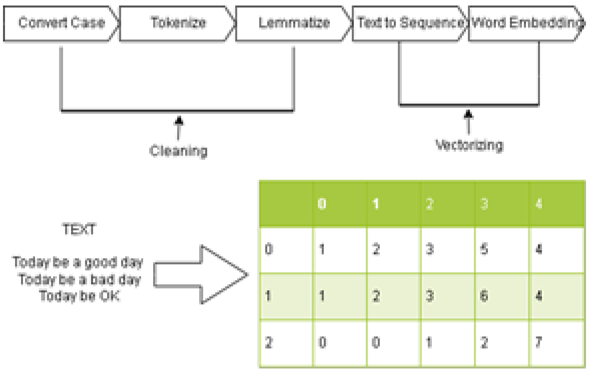
The Naive Bayes algorithm makes use of features that are easily understandable. Since the training phase is less computationally complex, this method is mainly used to perform sentiment analysis on large scale. It also has many limitations. It is highly depended on priors since it is a problematic classifier. Hence, training data should be representative all the time. Also it lacks in faulty inference on hidden data or text which is out of lexicon. The DTM model of Naive bayes possess features that are independent of each other, that means that lexical features in the DTM contribute in same proportion for all sentences irrespective of its relative text position. This results a training score of accuracy, testing score of accuracy or validation score of accuracy.

iii. Deep Learning

With Deep Learning (DL), data may be processed in a more complex manner. An LSTM model is a type of Recurrent Neural Network (RNN) for processing temporal data. We believe the words in the phrase are generated by the DL model's neural network design. With high-dimensional, sparse vectors, DL is computationally expensive. According to the Word embedding LSTM design, the model training must be represented as dense vectors. Which are the features retrieved from the original text. SVM for classification is used to discover a linear model of the following form.

$$y(x) = w^{T}x + b$$

equation 2Linear Model Discovery



	0	1	2	3	4
0	1	2	3	5	4
1	1	2	3	6	4
2	0	0	1	2	7

figure 5Word Embedding LSTM Data pre-processing weighted frequency count

One method is to convert each text into a series of integers, each of which represents a vocabulary term. Word embeddings must be used to map words that have similar usage or are similar to real number vectors. Open source pretrained models or custom neural network (unsupervised learning) models are used to extract embeddings. Current word embeddings can be taught alone or as an additional layer to the task's neural network model[1]. This is the method utilized since it produces embeddings that are specific to the user the data context as well as the goal. Furthermore, using Word embeddings with a sparse (hundreds of thousands of dimensions) DTM (Document Term Matrix) gives vectors with hundreds of dimensions while capturing semantic commonalities. One of the most important deep learning accomplishments for complex natural language for complex natural language processing(NLP) difficulties is Word Embedding[15].

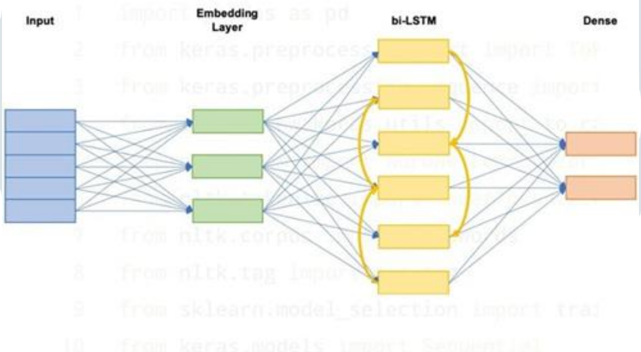


figure 6Word Embedding LSTM Architecture

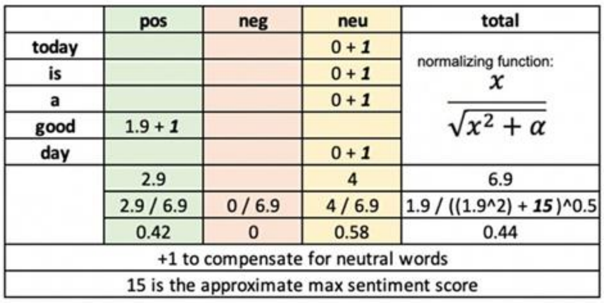
iv. *Pre-trainedandRulebasedVADERmodel*

Sentiment ratings are computed utilizing a broad Rule based tool which is the most in-demand approach. The VADER (Valence Aware Dictionary for sEntiment Reasoning) lexicon is a dictionary that assigns a sentiment score to each feature, which might be a word, expression, or acronym, ranging from the most negative to the most positive value. VADER makes use of a valence-score lexicon based that can detect sentiment strength on a given input text.

Because rule-based models are simple to understand and apply, they are a suitable option for emotional analysis. The disadvantage of rule-based models is that creating and validating lexicons takes time. Furthermore, this method evaluates individual words without taking into account the context in which they are used, which frequently results in errors, especially when sarcasm and irony are involved.

VADER calculates a composite score based on the sentiment intensity of the supplied text. It's calculated by summing the valence ratings of each lexical characteristic, modifying them according to the criteria, and then normalizing them to a range of -1 (the most extreme negative) to +1 (the most extreme positive) (most extreme positive) and the mid-point 0 represents neutral sentiment[5]. As a normalised, weighted composite score, the compound score is included here.

Further, Rule-based models are simple to understand and execute, making them an appealing option for emotive analysis. The problem with rule-based models is that creating and validating lexi-cons takes time. Furthermore, this method evaluates individual words while ignoring the context in which they are employed, which frequently results in inaccuracies, especially when it comes to sarcasm and irony[10].



	pos	neg	neu	total
today			0 + *1*	normalizing function: $\frac{x}{\sqrt{x^2+\alpha}}$
is			0 + *1*	
a			0 + *1*	
good	1.9 + *1*			
day			0 + *1*	
	2.9		4	6.9
	2.9 / 6.9	0 / 6.9	4 / 6.9	1.9 / ((1.9^2) + *15*)^0.5
	0.42	0	0.58	0.44
+1 to compensate for neutral words				
15 is the approximate max sentiment score				

figure 7Vadercalculations

VADER also relies on dictionary that map as words as shown in the fig 7. VADER is used to check polarity score after defining navies bayes algorithm. According to the formula where ‘x’ is sum of valence scores of constituent words, and ‘α’ is normalization constant (default value is 15). To check this we have to use Vader calculations to define output of particular text i.e ; positive, negative and neutral using normalization function. The reason behind this is that VADER is sensitive to both Polarity (whether the sentiment is positive or negative) and Intensity (how positive or negative is sentiment) of emotions this is provided by Valence Score to the word into consideration. It is a score assigned to the word under consideration by means of observation and experiences rather than pure logic.

IV.APPLICATIONS

Sentiment analysis a self regulated processofanalyzing the polarity of text and classify it into positive, negativeorneutral category.Some important applications of sentimentanalysis on business are listed below.

4.1Social Media Monitoring

In social media people share their thoughts, opinion and experiences of a particular service or product. Sentiment Analysis helps the businesses or influencers to perform in-depth analysis on the feedback from their clients and update their business strategies. With the help of Social Media Analysis tools we can quickly fetch thedataandanalyzeindividualresponsesas well as overall public sentiment on a particular topic. Looking at the customer feedback or reviews, businesses can update their marketing strategies that will improve their business.

4.2 Customer Support Management

Duetothe multiple numberofrequests,numerousthemes,and multiple divisions within a corporation, customer serviceadministrationcreatesmanyobstacles.

Sentimentanalysisisconcernedwith the understanding ofnaturallanguage,whichinvolves checking regular feedback and comments in order to comprehendclient requests. By moving sentiment to prioritize any urgentconcerns, you may automatically process customer supportconcerns, emails, online chats and phone calls. By analyzingclient phrases and words that include positive and negativesentiments,thisapplicationmightbeusedtosortthousandsofcustomersupportmessagesinashortamount of time[2].

4.3 Listen to Voice of Customer

Customer feedbacks can be gathered and evaluated from different sources like web, customer- satisfaction surveys, chats, call centers communications and emails. Sentiment analysishelps us to classifyandstructurethese collected raw data in order toidentify hidden patternsandrecurring contents with respect to a particular topic.As perthevoiceofcustomers,andknowingtheirwayofthinking, ithelpsustoocommunicatewiththecustomersinordertomakeitapersonalizedexperience [20].

4.4 Brand Monitoring and Reputation Management

Oneofthewell-known application ofsentimentanalysisinthebusinessworldisbrandmonitoring.Negativereviewsmaywreakhavoconyourbrand.Negativebrandcaptionsand mentions will be quickly alerted to you using sentimentanalysis technologies. It also comes with the ability to trackthe image and reputation of your brand. You can trackyour development inthe market with the helpof this. Youmay turn this data into actionable information by monitoringnews,blogs,andforumsforfeedbackonyourbrand.Youcan incorporate machine learning into this application to tracktrends and obtain results, allowing you to go from reactive toproactivemode [16].

4.4 Marketing and Competitor Review

Formarketandcompetitionresearch,usesentimentanalysisto see if you’re getting positive mentions from your competitors and to compare your marketing efforts. Analyze thepositive comments that your competitors use to communicatewiththeirconsumersandincludesomeofthesecommentsintoyourownproductbrandmessages,aswellassteerthetoneofyourcustomers’voicewhiledealingwithmarketconcerns[11].

IV.CONCLUSION

Sentiment analysis is to classify user’s opinion onvarioustopicsandarrangingresultintopositive,negativeandneutralandalsoitdeterminestheexperiencesof a user from textual content.Theviewsoftheuserscan be classified aspositive,negativeorneutral. It is important that data has to be fetched initially from various social media APIs to perform sentiment analysis. Since the collected data is not cleaned, data-pre-processing has to done as the second phase. In the third stage various computational techniques and models are used to classify the textual data based on polarity. Different techniques to perform sentiment analysis along with its benefits and drawbacks are discussed in this paper. Finally the outcome can be displayed as a visualization. This paper also discusses some important applications of sentiment analysis in the field of Business and marketing.

REFERENCES

[1] M. Rathi, A. Malik, D. Varshney, R. Sharma and S. Mendiratta, "Sentiment Analysis of Tweets Using Machine Learning Approach," 2018 Eleventh International Conference on Contemporary Computing (IC3), 2018, pp. 1-3, doi: 10.1109/IC3.2018.8530517.

[2] K. S. Naveenkumar, R. Vinayakumar and K. P. Soman, "Amrita-CEN-SentiDB: Twitter Dataset for Sentimental Analysis and Application of Classical Machine Learning and Deep Learning," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 2019, pp. 1522-1527, doi: 10.1109/ICCS45141.2019.9065337.

[3] A. Poornima and K. S. Priya, "A Comparative Sentiment Analysis Of Sentence Embedding Using Machine Learning Techniques," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 2020, pp. 493-496, doi: 10.1109/ICACCS48705.2020.9074312.

[4] B. K. Bhavitha, A. P. Rodrigues and N. N. Chiplunkar, "Comparative study of machine learning techniques in sentimental analysis," 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT), 2017, pp. 216-221, doi: 10.1109/ICICCT.2017.7975191.

[5] M. Patil and H. K. Chavan, "Event Based Sentiment Analysis of Twitter Data," 2018 Second International Conference on Computing Methodologies and Communication (ICCMC), 2018, pp. 1050-1054, doi: 10.1109/ICCMC.2018.8487531.

[6] A. K, K. P, L. Celestine S and V. V Kumar, "Naive Bayes Algorithm for Sentiment Analysis on Twitter," 2021 International Conference on System, Computation, Automation and Networking (ICSCAN), 2021, pp. 1-4, doi: 10.1109/ICSCAN53069.2021.9526473.

[7] S. Rana and A. Singh, "Comparative analysis of sentiment orientation using SVM and Naive Bayes techniques," 2016 2nd International Conference on Next Generation Computing Technologies (NGCT), 2016, pp. 106-111, doi: 10.1109/NGCT.2016.7877399.

[8] B. M, S. S, R. M, S. K. R and S. R, "A detailed study on sentimental analysis using Twitter data with an Improved deep learning model," 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2021, pp. 408-413, doi: 10.1109/I-SMAC52330.2021.9640850.

[9] R. B. Shamantha, S. M. Shetty and P. Rai, "Sentiment Analysis Using Machine Learning Classifiers: Evaluation of Performance," 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), 2019, pp. 21-25, doi: 10.1109/CCOMS.2019.8821650.

[10] V. Goel, A. K. Gupta and N. Kumar, "Sentiment Analysis of Multilingual Twitter Data using Natural Language Processing," 2018 8th International Conference on Communication Systems and Network Technologies (CSNT), 2018, pp. 208-212, doi: 10.1109/CSNT.2018.8820254.

[11] R. Mehra, M. K. Bedi, G. Singh, R. Arora, T. Bala and S. Saxena, "Sentimental analysis using fuzzy and naive bayes," 2017 International Conference on Computing Methodologies and Communication (ICCMC), 2017, pp. 945-950, doi: 10.1109/ICCMC.2017.8282607.

[12] J. Kaur and B. K. Sidhu, "Sentiment Analysis Based on Deep Learning Approaches," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), 2018, pp. 1496-1500, doi: 10.1109/ICCONS.2018.8662899.

[13] Y. Agarwal, R. Katarya and D. K. Sharma, "Deep Learning for Opinion Mining: A Systematic Survey," 2019 4th International Conference on Information Systems and Computer Networks (ISCON), 2019, pp. 782-788, doi: 10.1109/ISCON47742.2019.9036187.

[14] P. C. Shilpa, R. Shereen, S. Jacob and P. Vinod, "Sentiment Analysis Using Deep Learning," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), 2021, pp. 930-937, doi: 10.1109/ICICV50876.2021.9388382.

[15] Y. M. Wazery, H. S. Mohammed and E. H. Houssein, "Twitter Sentiment Analysis using Deep Neural Network," 2018 14th International Computer Engineering Conference (ICENCO), 2018, pp. 177-182, doi: 10.1109/ICENCO.2018.8636119.

[16] G. Kavitha, B. Saveen and N. Imtiaz, "Discovering Public Opinions by Performing Sentimental Analysis on Real Time Twitter Data," 2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET), 2018, pp. 1-4, doi: 10.1109/ICCSDET.2018.8821105.

[17] G. Subramaniam, R. Aswini, M. Ranjitha and P. K. Rajendran, "Survey on user emotion analysis using Twitter data," 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), 2017, pp. 998-1001, doi: 10.1109/ICECDS.2017.8389587.

[18] M. F. Çeliktuğ, "Twitter Sentiment Analysis, 3-Way Classification: Positive, Negative or Neutral?," 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 2098-2103, doi: 10.1109/BigData.2018.8621970.

[19] K. S. Naveenkumar, R. Vinayakumar and K. P. Soman, "Amrita-CEN-SentiDB 1: Improved Twitter Dataset for Sentimental Analysis and Application of Deep learning," 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2019, pp. 1-5, doi: 10.1109/ICCCNT45670.2019.8944758.

[20] G. Prema Arokia Mary, M. S. Hema, R. Maheshprabhu and M. Nageswara Guptha, "Sentimental Analysis of Twitter Data using Machine Learning Algorithms," 2021 International Conference on Forensics, Analytics, Big Data, Security (FABS), 2021, pp. 1-5, doi: 10.1109/FABS52071.2021.9702681.