

Twitter Sentiment Analysis Techniques

Mr. Lukesh Kadu¹, Dr. Manoj Deshpande² and Dr. Vijaykumar Pawar³

¹Shah and Anchor Kutchhi College of Engineering/Information Technology Department, Mumbai, India
Email: lukesh.kadu@sakec.ac.in

²⁻³A.C. Patil College of Engineering/Computer Engineering Department, Mumbai, India
Email: {mmdeshpande, vnpawar} @ acpce.ac.in

Abstract—Nowadays, social media networks are used by people incredibly frequently. They write articles, send private and public messages, pictures, and videos. In the realm of social networking, Tweeter is one of the most well-liked and often used social networks. Because of this, it's crucial to examine tweets in order to comprehend how people respond to a particular topic. For a variety of reasons, including corporate marketing, politics, studying public behavior, and data collection, it is essential to comprehend the emotion of tweets. Opinion study of Twitter text can help politicians and marketers better grasp how customers react to new product launches and marketing efforts. Particularly, Twitter is kind of the select few public sites where persons go for trustworthy data and is frequently used authoritatively by many people. This work uses a number of machine learning techniques to analyze tweets for sentiment using the terms in the tweets as input. The dataset that was used came from Kaggle. With layers for CNN, LSTM, BiLSTM, and embedding, a *deep learning* model has been created. As part of the analysis, Support Vector Machine (SVM), K Nearest Neighbor (KNN), Logistic Regression, Decision-Tree, Random Forest Method, and a deep learning architecture were used.

Index Terms— Machine learning (ML), Natural Language Processing (NLP), Convolution Neural Network (CNN), Deep Learning (DL).

I. INTRODUCTION

Sentiment analysis (SA) is a critical tool for capturing public sentiment from blog posts and social media in today's technologically advanced world. Web-based technologies are being used to spread a lot of information, and Internet users express their feelings via social media, blogs, and other media. Google is [1]. Sentiment analysis is the most crucial technique for identifying whether a certain opinion is good, bad, or unbiased comments on a variety of subjects, such as movie reviews, political viewpoints, worldwide pandemics, and other economic difficulties.

The techniques of locating and getting irrational data from a text, such as emotion, opinion, etc. is known as sentiment analysis. Numerous businesses across the globe are behind to identify normal ways for determining the sentiment of a given text since sentiment analysis is essential to identifying public input. identifies three categories for sentiment analysis techniques, including:

Decision trees continuously segment the data based on a specific parameter. Like decision nodes and leaves can describe the tree's structure.

The SVM is used for both regression and classification., a method for supervised machine learning. The best fit is classification, even if we also consider problems with regression. The SVM algorithm classifies the tweet text.

Positive, negative, or neutral emotions are classified according to their polarity (neutral). It's critical to draw attention to the following levels on which sentiment analysis can be performed

- Document-level sentiment classification: A document can be classed entirely as "good" "bad" or "impartial" at this level.
- Sentiment categorization at the sentence level: Every tweet is categorised as "good" "bad" or "impartial" at this level.[2]

At this level, sentences and documents can be classified as "good" "bad" or "impartial" based on certain elements of the language and archives." The phrase "perspective-level assessment grouping" is another name for this level. The benefits of deep learning models include automatic feature extraction, simple calculation due to the use of accelerated hardware, and the best performance even with enormous amounts of data. Deep learning is a technique that is being employed at several places. It is a component of a neural network that has numerous hidden layers and replicates the mechanism by which human brain neurons function to produce precise results.

As a result, the sentimentality of the tweets is improved. This system's suggested approach employs sentiment analysis on Twitter tweets. The info obtained via the Twitter API is set. Pre-processing procedures are used to sanitise the words, including stemming, removing punctuation, stop words, and white spaces, as well as changing all of the tweets' case. The sentence is then tokenized. In both Korean and English, the data set comprises of numerous training value and numerous testing value. The data set is then subjected to FNN.

II. RELATED WORK

A well-known use of natural language processing is Twitter sentiment analysis. The data used in the research and the kinds of approaches created to do this objective vary. This study examines the relationship between views and opinions and how these things might be expressed in natural language about an occurrence or event. The current situation demonstrates that sentiment analysis has advanced to the point where it can now deal with behavior and emotions for a variety of subjects and languages in addition to positive and negative sentiment. In the study of sentiment analysis, researchers employ several methods for foretelling social opinion and emotion through language and text.

Tweets are utilized as the raw data, as recommended by T. Jayasankar, Avudaiappan.T, Jenifer, Sisay tumsa, and others. A secret token is used to gather the tweets over the Twitter API.

The noise in the words is then reduced using a text mining. Program during preprocessing. Each pre-processed sample has its calculated score.

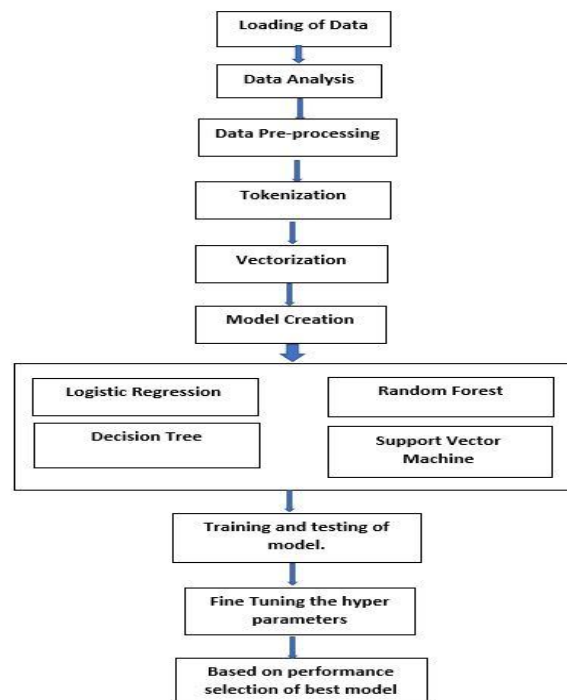


Figure.1. Proposed work Flowchart [3]

The score is given out as 1, -1, or 0 for neutral and negative tweets, respectively. The results of the pre-processing of the tweets are recorded in CSV format for further processing [4]. The ratio of the train and test data sets for building the classification model is 60:40. Following categorization, it is discovered that the CNN Network is not formed to determine the likelihood of the tweets. To improve on the holdout approach, the system makes use of the K-fold cross validation method. As a last step, a hybrid technique is used to forecast the likelihood of positive tweets in relation to the given tweets. Comparing the results of this system with others, to other method performs better than CNN and simple RNN for sentiment classification.

Mohammad Zubair Khan and Abdullah Alsaedi make a proposal. Sentiment analysis seeks to pinpoint the feelings that underlie opinions in texts on a range of subjects. Academics studying sentiment analysis have recently concentrated on assessing opinions on a range of topics, such as movies, commercial items, and common societal problems. Customers can express their ideas on Twitter, an incredibly popular microblog. Researchers can examine the results of several sentiment analyses performed on Twitter data [2].

Hong Soon Goo and Adyan Marendra Ramadhani deliver Twitter is an SNS that collects a lot of data from user postings. Because of this sizeable data set, text mining and sentiment analysis studies may be possible. However, managing such a large amount of unstructured data is a challenging undertaking, necessitating the use of machine learning which employs a feed-forward numerous hidden layers [5].

According to Vedant Pandya, Anuja Somthakar, and Shivanshu S Shrivastava, Twitter is incredibly well-liked, with a significant user base and a commanding global presence, Twitter is immensely well-liked. Twitter is a platform used by people from many walks of life to voice their ideas on a variety of subjects. So, businesses can gain useful information about public opinions by using Sentiment Analysis techniques on tweets, which can aid them in making better judgements. The 1.6 million tweets in this dataset were divided into positive and negative categories using machine learning and deep learning models. The accuracy, F1-Score, recall, and precision criteria were used to compare the performance of various models [6].

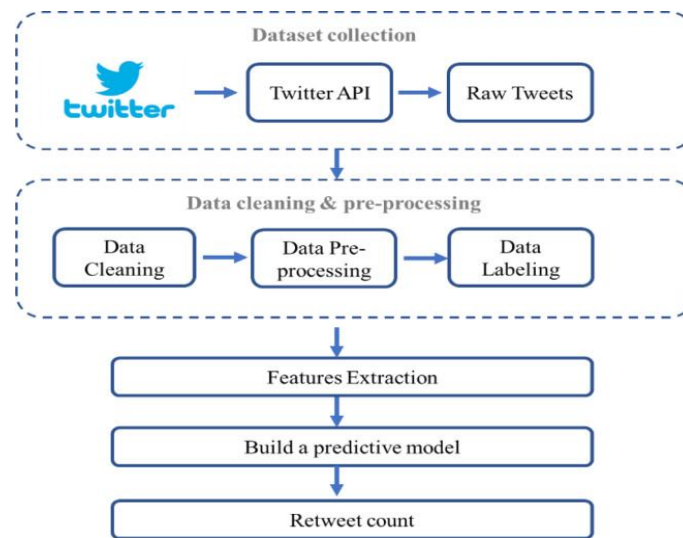
The invention was made by Poornima A, Nataraj N, Nithya R, Nirmala D, and Divya P Sentimental analysis is the computational examination of individuals' attitudes, beliefs, and feelings about various objects, people, occasions, events, or topics. The accuracy of sensation exploration has been studied extensively, with studies ranging from straightforward more direct models to intricate Deep Neural Network models. Deep Learning is now thought to be the best model in its field because of its recent success in opinion analysis. Due to its complex layout, several dialects, and lack of resources, Twitter information presents a number of challenges. Even if the Twitter feeling examination's accuracy has improved thanks to the new profound learning model, there is still room for improvement. This compelled us to improve the accuracy of the Twitter opinion analysis, look into several methods that poor people have applied to information from Twitter. [7]

Tweets that were all about training a dataset. 1600000.processed.noemotico. The tweets were categorized as good, negative, or neutral using sentiment analysis. In order to examine the performance of Several supervised learning methods, including Naive Bayes, Support Vector Machines, Decision Trees, and Bagging, were taught. Cross-validation revealed that Maximum Entropy produces the best outcomes. highlights the many classifications that product criticisms can be given. After carefully preprocessing the data and extracting features using Bigrams and unigrams, they constructed and compared models such as SVMs, Naive Bayes, Decision Tree, Random Forest, and XG Boost.

III. METHODOLOGY

Positive and negative tweets are used to train a sentiment analysis model. As seen in fig.2, Before being provided to various models for sentiment analysis, tweets are loaded, pre-processed, displayed, and transformed into features.

1. **Loading of Datasets** The Kaggle-available training.1600000.processed.noemotico dataset was used to train the model. It is one of the biggest database for sentiment analysis utilizing Twitter data exists here..
2. **Dataset Preprocessing**
Dataset Preprocessing Kaggle datasets originate from a wide range of sources and are frequently in a wide range of formats. This makes cleaning our raw data a crucial step in the dataset preparation process. Yet, as textual content commonly uses repeated or redundant words, cleaning is a difficult task. Before training the model, we will preprocess the dataset in a number of ways, including:
 - Removing stop words.
 - Removing emojis.
 - Removing of mentions.



- Removal of numbers.
- Removal of whitespaces.
- Removal of duplicated rows.
- Elimination of odd columns.
- Converting the text document to lowercase for better generalization.
- Correcting punctuation (to reduce unnecessary noise from the dataset)
- Eliminating repeated characters from the text as well as URLs and hyperlinks that don't provide a functional purpose.
- High degree of knowledge of a language.

3. Dataset Visualization



The Word Cloud created from the dataset's negative tweets is shown in Fig. 4, and it demonstrates that negative tweets frequently contain the phrases "don't," "miss," and "awful."

$$\text{PRECISION} = \text{TP} / (\text{TP} + \text{FP}) \quad (1)$$

$$\text{RECALL} = \text{TP} / (\text{TP} + \text{FN}) \quad (2)$$

$$\text{F1} = (2 * \text{PRECISION} * \text{RECALL}) / (\text{PRECISION} + \text{RECALL})$$

$$\text{ACCURACY} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FN} + \text{TN} + \text{FP}) \quad (3)$$

$$\text{SPECIFICITY} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (4)$$

6. Deep Learning models used for training

A. Recurrent Neural Network

Recurrent neural networks (RNNs) are a subset of networks useful for researching how future data, such as natural language processing, is represented. It generates an objective function that is dependent on past state outputs or hidden states as well as the current input. Earlier state input determines the earlier state output in this case. According to the prior input and output, the current state output is determined [6].

TABLE I. RESULT FOR RNN

Model	Metrics			
Recurrent Neural Network	Accuracy	F1	Recall	Precision
	78.15%	78.04%	77.76%	78.36%

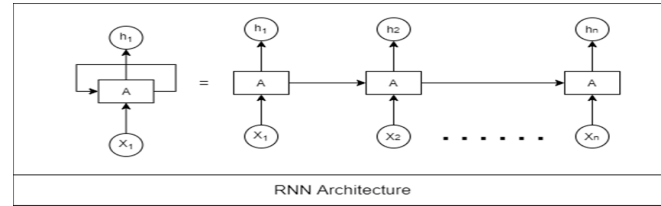


Figure 5. Architecture of RNN

B. Long Short-Term Memory

The cell state and different gates make up the bulk of an LSTM. The information is transferred along the sequence chain via the cell state. The information necessary for the full processing of the entire sequence is stored in the cell state, which functions as a memory. Here, the RNN's short-term memory problem is solved, allowing even pertinent data from the earliest time steps to influence data from later time steps. As a result, during the training phase, relevant information is added and irrelevant information is subtracted using gates in the cell state.[9-12]

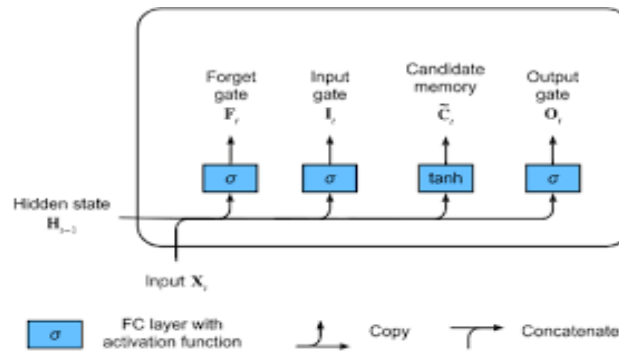


Figure 6. Architecture of LSTM

TABLE II. Result for LSTM

Model	Metrics			
	Accuracy	F1	Recall	Precision
Long Short-Term Memory	80.21%	80.3%	79.82%	80.83%

C. Bidirectional LSTM

One RNN approach to enhance LSTM, which has weaknesses in text sequence characteristics, is Bi-LSTM. It performs sequential modelling task more well than LSTM. In LSTM, information flows from back to front. Bi-LSTM is a pioneer in sentiment classification thanks to its structure, which allows it to better understand the context. Bi-architecture LSTM's is depicted in Figure 4. Using two different approaches, the input data from the previous and following sequences are preserved in Bi-LSTM, in contrast to the conventional RNN model, which requires decay to accommodate future data.

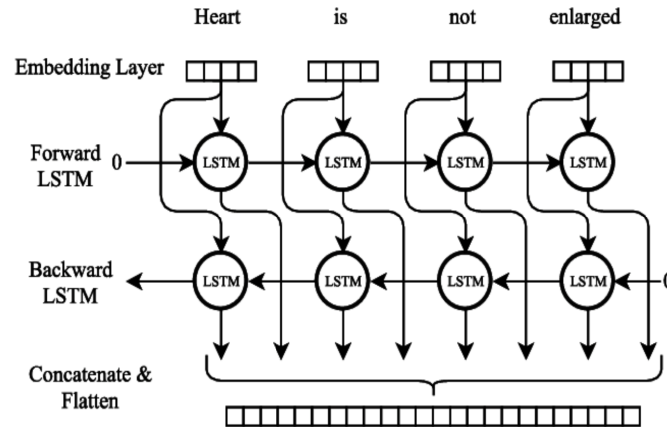


Figure 7. Architecture of BiLSTM

TABLE III. RESULT FOR BI LSTM

Model	Metrics			
	Accuracy	F1	Recall	Precision
Bidirectional LSTM	81.33%	81.54%	80.99%	82.13%

IV. EXPERIMENTAL RESULT AND DISCUSSION

It comprises the outcomes and functionality of each model that was employed for the sentimental analysis of Twitter tweets. Datasets with different levels of open access and authentication were available on Kaggle, where the dataset for this challenge was obtained. [9]

For document-level sentiment categorization, the three deep learning architectures basic RNN, LSTM, and BiLSTM are contrasted. The training and testing accuracy attained for all three networks is depicted in the figures below. It is evident from the data that BiLSTM per accuracy. This is due to the fact that the LSTM network outperforms both CNN and basic RNN models in terms of word sequence memory.

TABLE IV: RESULTS OF DEEP LEARNING MODEL

Model	Metric			
	Accuracy	F1	Recall	Precision
RNN	78.15%	78.04%	77.76%	78.36%
LSTM	80.21%	80.31%	79.82%	80.83%
BI LSTM	81.33%	81.54%	80.99%	82.13%

In the table above, BiLSTM algorithm accuracy is 81.33%, whereas RNN accuracy is 78.15%. In comparison to RNN, which has a recall of 77.76% and a precision of 78/36%, BiLSTM has a recall of 80.99% and a precision of 82.13%. The machine learning model yields the result as shown below.

TABLE V: RESULT OF MACHINE LEARNING MODEL [8]

Model Id	Model Name	Accuracy	F1-score (class 0)	F1-score (class 1)	AUC Score	Trainig execution time in seconds	Testing execution time in seconds
1	Bernoulli Naive Bayes (BNB)	73%	73%	74%	73%	0.44	0.77
2	Support Vector Machine (SVM)	74%	73%	75%	74%	34.38	0.63
3	Logistic Regression (LR)	74%	74%	75%	74%	36.13	0.67
4	K-nearest neighbors (KNN)	60%	NaN	NaN	61%	0.16	NaN
5	Decision Tree (DT)	69%	69%	69%	69%	31.84	0.76

When we trained the models in terms of accuracy, deep learning algorithms outperform machine learning algorithms. F1 Score, Recall, Precision, and training execution time. Machine learning algorithms like Naïve Bayes, Logistic Regression, SVM, KNN, Decision tree and deep learning algorithms like RNN.LSTM, BILSTM are trained by using training. 1600000.processed.noemotico dataset.

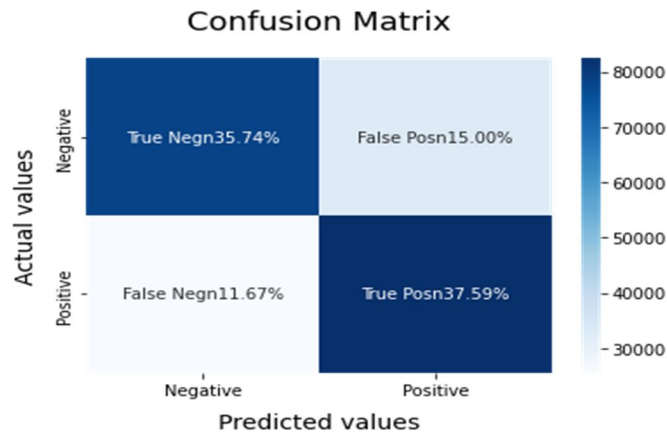


Figure 8. Conflict matrix for Naïve Bayes

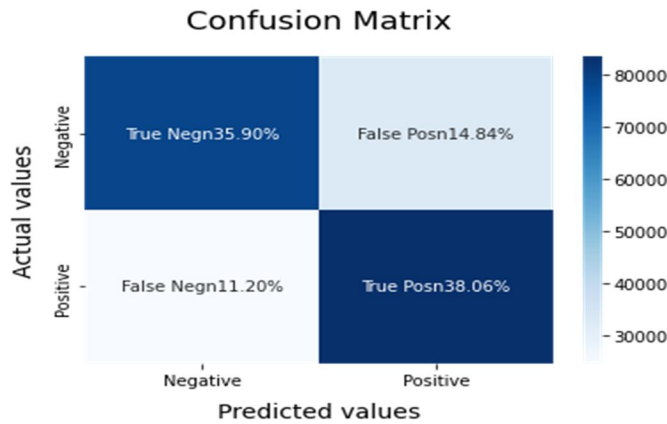


Figure 9. Conflict matrix for SVM

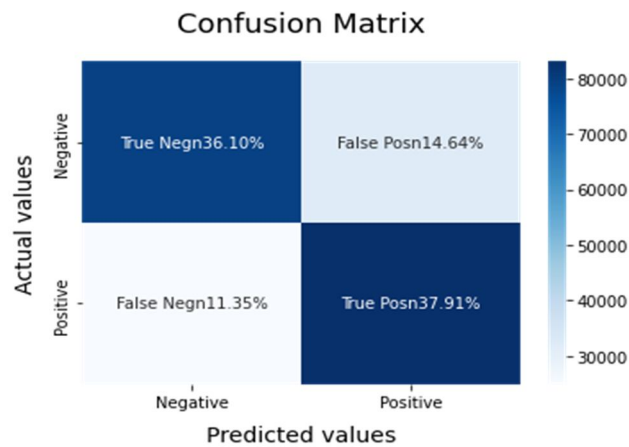


Figure 10. Conflict matrix for logistic regression

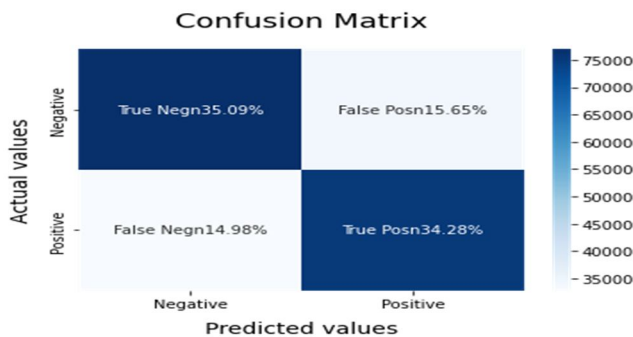


Figure 11. Conflict matrix for Decision Tree

Execution time: Bernoulli Naive Bayes performs better with a higher accuracy rating when comparing the execution times of different models.

Model accuracy: With an accuracy of 74%, the logistic regression & Support Vector Machine model outperforms the majority of the other models

V. CONCLUSIONS

A simple analysis of the type of information shared on Twitter, one of the This work applies a range of deep learning and machine learning models to sentiment analysis on tweets. When various types of machine learning models were trained on the dataset and evaluated for performance, they outperformed Naive Bayes with an accuracy of 76.53%.. Both models performed admirably, but DL models were superior. RNN, LSTM, and Bi-LSTM, three deep learning models with sophisticated architecture and context understanding, performed wonderfully; Bi-LSTM was the best performer, with an accuracy of 81.33%. DL algorithms required a lot of computing power and time to train, but they excelled in sentiment analysis. Future sentiment categorization analysis may examine the effectiveness of various deep learning models.

REFERENCES

- [1] Tajebe Tsega Mengistie & Deepak Kumar, "Deep Learning Based Sentiment Analysis On COVID-19 Public Reviews": ICAIIC | 978-1-7281-7638-3/20/\$31.00 ©2021 IEEE | DOI: 10.1109/ICAIIIC51459.2021.9415191
- [2] Abdullah Alsaedi & Mohammad Zubair Khan. "A Study on Sentiment Analysis Techniques of Twitter Data" (IJACSA), Vol. 10, No. 2, 2019,
- [3] Sai Kumar T S , Arunaggi Pandian K, Thabasum Aara S, Nagendra Pandian K "A Reliable Technique for Sentiment Analysis on Tweets via Machine Learning and BERT" (IANCON) | 978-1-7281-8402-9/21/\$31.00 ©2021 IEEE | DOI: 10.1109/ASIANCON51346.2021.954501
- [4] Avudaiappan.T, Jenifer, Sisay tumsa, Subashree, T.Jayasankar, "Twitter Sentimental Analysis Using Neural Network", IJSTR©2020 VOLUME 9, Issue 02, February 2020 ISSN 2277-8616

- [5] Adyan Marendra Ramadhani and Hong Soon Goo, "Twitter Sentiment Analysis using Deep Learning Methods." Engineering Seminar (InAES), 2017 7th International Annual.
- [6] Vedant Pandya, Anuja Somthankar, Shivanshu S Shrivastava & Dr. Megharani Patil "Twitter Sentiment Analysis using Machine Learning and Deep Learning Techniques" 2021@IEEE | DOI: 10.1109/C21454156.2021.968924
- [7] Poornima A , Nataraj N , Nithya R , Nirmala D , Divya P "Sentiment Analysis of Tweets in Twitter Using CNN" IEEE Xplore, I-SMAC 2021: CFP21OSV-ART; ISBN: 978-1-6654-2642-8
- [8] <https://www.kaggle.com/code/mohamedisbaine/tweeter-sentiment-analysis-nlp>
- [9] Bhavani M, Shrijeeth S, Rohit M, Sanjeev Krishnana R, Sharveshwaran R "A detailed study on sentimental analysis using Twitter data with an Improved deep learning model" IEEE Xplore, I-SMAC 2021: CFP21OSV-ART; ISBN: 978-1-6654-2642-8
- [10] Irfan Aygün , Buket Kaya , and Mehmet Kaya, "Aspect Based Twitter Sentiment Analysis on Vaccination and Vaccine Types in COVID-19 Pandemic With Deep Learning", IEEE Journal of Biomedical and Health Informatics, Vol. 26, No. 5, May 2022.
- [11] S. Singh, A. Dixit, and G. Joshi, "'Is compulsive social media use amid COVID-19 pandemic addictive behavior or coping mechanism?", Asian Journal of Psychiatry, vol. 54, p. 102290, 2020
- [12] S. Tiwari, A. Verma, P. Garg and D. Bansal, "Social Media Sentiment Analysis on Twitter Datasets," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 2020, pp. 925-927, doi: 10.1109/ICACCS48705.2020.9074208.



Mr. Lukesh R. Kadu working as a assistant professor at Shah and Anchor Kutchhi College of Engineering, Chembur ,Mumbai and research scholar at AC Patil college of engineering, Kharghar, Navi Mumbai. His expertise in software engineering and domain of work is deep learning and machine learning.



Dr. Manoj M. Deshpande working as a professor at AC Patil College of Engineering, Kharghar, Navi Mumbai and research Guide at AC Patil college of engineering. He received his Ph.D. From IIT



Dr. V N. Pawar working as a professor & Principal at AC Patil College of Engineering, Kharghar, Navi Mumbai and His area of specialization is Machine Vision, Robotics and Embedded System Design