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

Sentiment Analysis Of Text Using MachineLearning

[1]Anuradha Tanaji Khade, [2] Shweta Sambhaji Ghorpade, [3] Rohit Nandkumar Bhosale, [4]Rushikesh Ramesh Kesare, [5] Mrs.Hafsa Shoyeb Majgoankar

[1,2,3,4] UG Students, [5]Assistant Professor
[1,2,3,4,5] Department of Computer Science and Engineering
[1,2,3,4,5] Nanasaheb Mahadik College of Engineering, Peth, Maharashtra, India

Abstract : In recent days, invention of new platforms in social media has given lot of boost to the business development. In the business process social media is playing an important role as a deciding factor for success or failure of a business in a growing economy of the country. One such platformwhich helps people to understand and gauge the business prospectus is twitter. In this paper we are addressing the problem of sentiment analysis in twitter; which mainly deals with classifying tweets according to the sentiment expressed in them: positive, negative or neutral. Twitter is an online micro-blogging and social-networking platform which allows users to write short status updates of maximum length 140 characters. Analyzing the public sentiment is important for many applications such as firms trying tofind out the response of their products in the market, predicting political elections and predicting socioeconomic phenomena like stock exchange. The aim here isto develop a functional classifierfor accurate and automatic sentiment classification of an unknown tweet stream.

IndexTerms **- : Machine learning, Deep learning, Twitter datasets, Positive words, Negative words .**

I. INTRODUCTION

The popularity of social networking sites like Twitter is growing along with social awareness. Twitter is a significant and wellliked social media platform where anyone can tweet about any event. People are able to openly share their thoughts, beliefs, and emotions on an open platform. People have twitter accounts due to cheaper internet costs, more affordable portable devices, and more social importance. The majority of them tweet about various occurrences. People in the social networking era use Twitter to convey their thoughts and emotions. Thus, Twitter is extremely data-rich. Every tweet's length is known to us. Opinion mining, often known as sentiment analysis, is nothing more than the examination of attitudes or feelings in text data. Sentiment analysis identifies each person's opinion or sentiment with regard to a particular occurrence. We must submit aprovide a distilled version of the document's opinion. One newer and difficult topic of research is Twitter sentiment analysis. It is helpful to determine feelings or opinions of individuals regarding specific events because social media sites like twitter offer a significant volume of text sentiment data in the form of tweets. For reviews of films, products, customer service, opinions about any event, etc., sentiment analysis or opinion mining is useful. This aids in determining if a particular commodity or service is preferred or not. It is also helpful to determine people's opinions about any situation or individual, and it may determine whether a text is good, negative, or neutral. Text can be classified into various sentiments using a type of text analysis called sentiment analysis. The technique of identifying the sentiment or emotional tone expressed in a piece of text, such as positive, negative, or neutral, is known as sentiment analysis, sometimes known as opinion mining. Because machine learning techniques can automatically identify patterns and characteristics from data, they are frequently .

II. LITERATURE REVIEW

For the sentiment classification of movie reviews, a hybrid classification technique has been applied. To evaluate performance based on accuracy, various feature sets and classification algorithms, such as Naive Bayes and Genetic algorithms, have been integrated. The results of the research demonstrate that hybrid NB-GA is more efficient and effective than base classifier, and that GA is more efficient than NB when comparing the two. [1]

Text mining must also consider a document's polarity. The topic of future engineering using tree kernels has been covered by . Compared to other procedures, this one produces better results. The author of the paper defines two classification models, a 2-way classification and a 3-way classification. In a two-way classification, emotions can be positive or negative, and in a three-way classification, they can be positive, negative, or natural. The author considers the tree kernel approach for tweet representation. The most accurate and best

feature-based model was a tree kernel model. The experiment outperforms the unigram model by 4%.[2]

For cascaded categorization, a hierarchical approach to sentiment analysis can be applied To create a hierarchical model, the author cascaded three classifications: positive against negative, polar versus non-polar, and objective versus subjective. This approach was evaluated in comparison to a 4-way classification model (objective, neutral, positive, and negative). The comparison's results reveal that the hierarchical model performs better than the 4-way categorization approach. [3]

The author has created a domain-specific feature-based model for movie reviews. In this case, an aspect-based technique is utilized to analyses text movie reviews and assign a sentiment label to them.The sentiment score for a particular movie is then calculated by averaging each factor over numerous evaluations. Author employs technique based on SentiWordNet for feature extraction and sentiment analysis at the document level. Alchemy API results are compared to the algorithmic result. Thecomparison reveals that the result of the feature-based model is superior than the Alchemy API technique. Simply put, aspect-based sentiment results are superior to document-based results. [4]

The author has gathered a sizable corpus of close to 300000 tweets for sentiment analysis and opinion mining. It is possible to classify tweets as positive, negative, or neutral using a sentiment classifier model. Positive emotions, such as joy, happiness, or amusement; Negative emotions, such as sadness, wrath, or disappointment; and Neutral-text that doesn't include any emotions-were the three categories into which the collected corpus was split in this technique. For POS-tagging to distribute emotions, Tree Tagger is utilized.[5]

Consumer marketing information is used to gather opinions about products and to predict the future. Since there is a vast amount of consumer review data, the author employs the Hadoop environment for sentiment analysis. Hadoop clusters were built as part of an experiment for data analysis. Positive, negative, and neutral tweets were classified [6].

The HIVE and FLUME tools from Hadoop are also used to analyse twitter data. Data is extracted using the FLUME programme and stored in HDFS format. Data from HDFS-type storage is extracted and analysed using the HIVE programme. HIVE tool aids in the examination various[7]

According on the search terms used in tweets about customer reviews, Twitter data is also automatically categorised into good, negative, and neutral categories. The author of the paper employs the tree kernel and Parts of Speech (POS) polarity techniques. The two sorts of resources used in research include a manual dictionary of emotions and a lexicon compiled from the internet. The author employed various feature extraction and classification algorithms.[8]

A thorough ordinal regression-based machine learning algorithmic analysis of tweet sentiment was sought after by Saad and Yang in 2019. Pre-processing tweets is a stage in the suggested methodology, and the feature extraction model was used to create an effective feature. Techniques like SVR, RF, Multinomial Logistic Regression (SoftMax), and DTs were utilised to categorise the sentiment analysis. Additionally, the proposed model was examined using data from Twitter.[9]

AFramework for Cybercrime Prediction on Twitter Tweets Using Text- Based Machine Learning Algorithm: With the rise in popularity, social media platforms such as Twitter are more recognized for calling tweets to constructuser networks that can communicate. Because of the increasing number of Twitter users and the expansion in cybercrime rates. Cybercrime is a crime perpetrated using technology.It is feasible that the social networking platform might be utilized to commit a crime. The researchers focused on cyberbullying and cyberthreat, given the severity of cybercrime's effects on victims and given the devastating effects of cybercrime on victims, it is critical to devise effective methods for predicting and preventing it. ”2022[10].

III .PROPOSED SYSTEM:

3.1 System Architecture:

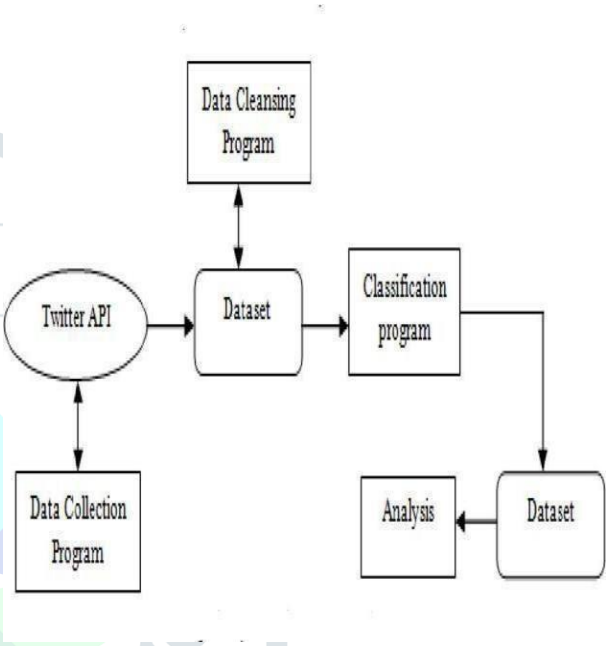

Figure 1 : System Architecture

Step-1 1]Installation of the needed software authentication of twitters data. The main installation software’s include tweepy, text blob, nltk etc, Authentication involves different steps step1: visit the twitter website and click the button ‘create new app’.

Step-2 fill the details in the form provided and submit Step-3 It will be redirected to the app page where the “‘consumer keys’, ‘consumer access’, access token’ and ‘access token secret’ “that is needed to access the twitter data will be present.Step-4 implement in python. There are different sources for storing the data taken from the twitter. They are like MongoDB , open source document storage database and is the go-to “No SQL” database. It makes working with a database feel like working with JavaScript. PyMongo, a Python wrapper for interfacing with a MongoDB instance. This library lets you connect your Python scripts with your database and read/insert records. This is an example of the data thatis been extracted from the twitter on the topic computer using python code.

Once the data is collected from the twitter the next step is preprocessing that is implementedin python. There are

several steps involved in the preprocessing stage. They are Converting all uppercase letters to lowercase. Tokenization generally done by installing the NLP package. It generally means removal of hash tags, numbers (1, 2, 3 etc.,), URL's and targets (@). Once tokenization is over we move to the next step of preprocessing Removal of non-English words Twitter generally supports more than 60 languages. But our project mainly involves English tweets; hence we remove the non-English words.Emoticon replacements Emoticons are very important in determining the sentiment. Sothe emoticons are replaced by their polarity by seeing the emoticon dictionary.

Removal of stop words Stop words play a negative role in sentimental analysis, so it is important to be removed. They occur both I negative and positive tweets. A list of stop wordslike he, she, at, on, a, the, etc. are created and ignored. Once the above four steps are over we move to the next main method called feature extraction

3.2 .Data Flow Diagram:

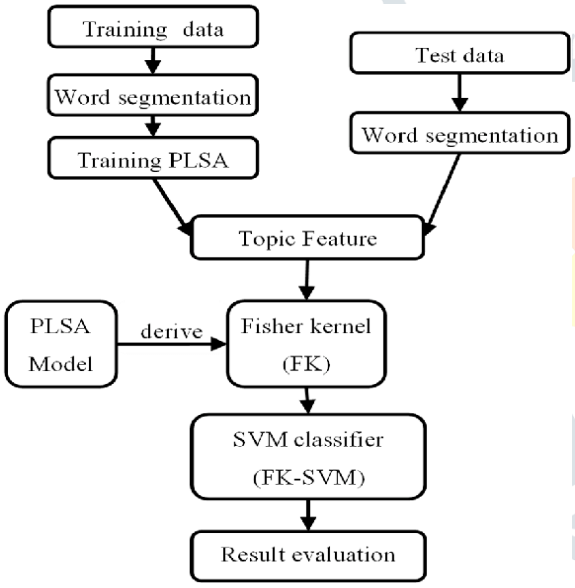

Figure 2 : Data Flow Diagram

A data flow diagram (DFD) is a graphical representation of the flow of data through a system.It is used to model the flow of information in a system and to identify the transformations thatdata undergo as it moves through the system.DFDs can be used to model a wide range of systems, including information systems, business processes, and software systems. They are useful for visualizing and analyzing the flow of datathrough a system, Identifying bottlenecks or inefficiencies, and communicating the design of a system to others.There are several different notations that can be used to create DFDs, including the Gane-Sarson notation and the Yourdon-DeMarco notation.

The specific notation used will depend onthe preferences and conventions of the person creating the DFD.

IV .METHODOLOGY:

These systems don't depend on manually crafted rules, but on machine learning Techniques, like classification.Classification,which is employed for sentimentanalysis, is automatic system that must be fed sample text before returning a category, e.g. positive, negative, or neutral. Urgent issues will often arise, and they must be restrainedimmediately. A complaint on Twitter, for instance, could quickly escalate into a PR crisis if it goes viral. While it'd be difficult for your team to spot a crisis before it happens, it's very easy for machine learning ools to identify these situations in real-time.Patterns are often extracted from analyzing the frequency distribution of those parts of speech (either individually or collectively with some other parts of speech) during a particular class of labeled tweets.

4.1 ALGORITHMS

1. Long short term memory)

Recurrent neural networks (RNNs) are a form of Artificial Neural networks that can memorize arbitrary-length sequences of input patterns by capturing connections between sequential data types. However, due to stochastic gradients' failure, RNNs are unable to detect long-term dependencies in lengthy sequences. Several novel RNN models, notably LSTM, were proposedto address this issue. LSTM networks are RNN extensions designed to learn sequential(temporal) data and their long-term connections more precisely than standard RNNs. They arecommonly used in deep learning applications such as stock forecasting, speech recognition, natural language processing, etc.

Sentiment analysis is a potent tool with varied applications across industries. It is helpful for social media and brand monitoring, customer support and feedback analysis, market research, etc. A new product's target audience or demographics can be identified by performing sentiment analysis on initial customer feedback received

2. Logistic Regression

Given a tweet, or some text, we can represent it as avector of dimension V, where V correspondsto our vocabulary size. For example: If you had the tweet"I am learning sentiment analysis", then you would put a 1 in the corresponding index for any word in the tweet, and a 0 otherwise.As we can see, as V getslarger, the vector becomes more sparse. Furthermore,we end up havingmany more features and end up training θ V parameters. This could result in larger training timeand large prediction time. Hence, we will extract frequencies of every word and making a frequency dictionary The idea here is to divide the training set into positive and negative tweets.Count

all the words and make a python dictionary of their frequencies in positive and negative tweets.For every tweet make a vector of bias unit, sum of all the positive frequencies(words from positivetweets) of all the words and also their negative frequencies. We will go into detail regarding this in further paragraphs.

3. RNN(Recurrent neural network)

Recurrent Neural Networks (RNN) are to the rescue when the sequence of information needed to be captured (another use case may include Time Series, next word prediction, etc.). Due to its internal memory factor, it remembers past sequences along with current input which makes itcapable to capture context rather than just individual words. They are the most basic form of Recurrent Neural Networks that tries to memorize sequential information. However, they have the native problems of Exploding and Vanishing gradients. For a detailed understanding of how RNNs works and its limitations The vanillaform of RNN gave us a Test Accuracy of **64.95%**. Limitations of Simple RNN are it is unable to handle long sentences well because of its vanishing gradient problems.

4. Natural Language Processing

An interdisciplinary topic within linguistics, computer science, and artificial intelligence called "natural language processing" studies how computers and human language interact, with afocus on how to programmed computers to handle and analyze massive amounts of natural language data. Artificial intelligence (AI) includes the field of natural language processing (NLP). In many different commercial disciplines and places, personal assistants employ this technology.

This technology processes the user's speech in accordance with its proper understanding by dissecting it. Due to the fact that it is a very currentand successful strategy, there is a huge demand forit right now. In the nascent field of natural language processing, advancements like smart device interoperability and interactive human conversations have already been made pos

V. EXPERIMENTAL WORK:

PRECISION RECALL:-

While building any machine learning model, the first thing that comes to our mind is how we can build an accurate & 'good fit' model and what the challenges are that will come during the entire procedure. Precision and Recall are the two most important but confusing concepts in Machine Learning. Precision and recall are performance metrics used for pattern recognition and classification in machine learning. These concepts are essential to build a perfect machine learning model which gives more precise and accurate results. Some of the models in machine learning require more precision and some model requires more recall. So, it is important to know the balance between Precision and recall or, simply, precision-recall trade-off. 4.4.1 Precision and Recall in Machine Learning:- This matrix consists of 4 main elements that show different metrics to count a number of correct and incorrect predictions. Each element has two words either as follows: There are four metrics combinations in the confusion matrix, which are as follows: True Positive: This combination tells us how many times a model correctly classifies a positive sample as Positive? False Negative: This combination tells us howmany times a model incorrectly classifies a positive sampleas Negative? False Positive: This combination tells us howmany times a model incorrectly classifies a negative sample as Positive? True Negative: This combination tells us how many times a model correctly classifies a negative sample as Negative? 4.5.2 Confusing Matrix Precision is defined as the ratio of correctly classified positive samples(True Positive) to a total number of classified positive samples (either correctly or incorrectly). Precision = True Positive/True Positive + False Positive Precision =TP/TP+FP TP- True Positive FP- False Positive The precision of a machine learning model will be low when the value of; TP+FP (denominator) > TP (Numerator) The precision of the machine learning model will be high whenValue of; TP (Numerator) > TP+FP (denominator)

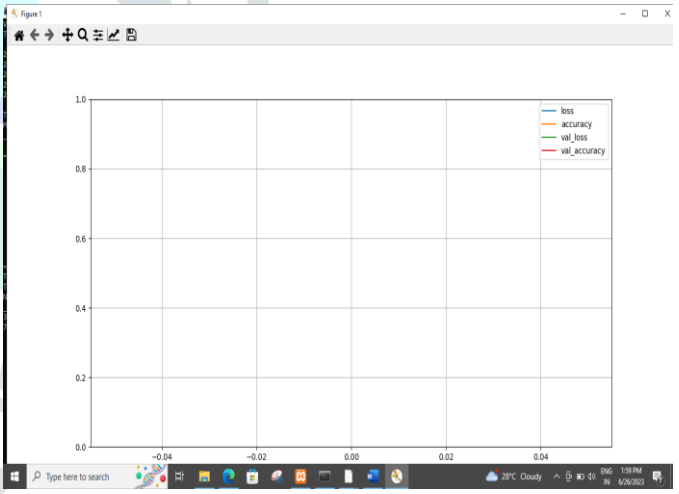

Figure3.AccuracyGraph

VI RESULTS AND DISCUSSION

6.1 Implementation and Output

This chapter discusses about the implementation, deployment, and result of the entire application after being developed. The implementation process is must need a method to carryout, execute the project after the system design. The system being implemented into a real prototype or integrate software-based service for the end-user. After implementation, the system testing

is executed to test the whole system for the functionality and credibility of the system being developed. In this process, the algorithm or technique being applied along with the development of the application.

6.2 Deployment and Configuration

In this stage, the deployment takes place on deploy the system requirements to enable development of this project. The hardware requirement being setup and testing either it suitable and compatible with the project requirement. The process conducted by allowing thevirtualization for Intel Core i5 which allows android is running with better graphics and virtualize the emulator. The process deployment of AWS Cloud as a web server, python machine learning and MySQL that need to configure and deploy to develop an stand alone software. Configuration and deployment being implement into Android Studios byusing python language and connect with jupyter notebook for comparison process to authenticate the user. The process conducted involving software and hardware requirement based on system design to ensure all meet the expectation. 5.1.2 Interface The interfaces are a central part of stand alone software where by shows the flow of interfacesof sentiment analysis of twitter text.

6.3 RESULTS:

6.3.1 : Main GUI

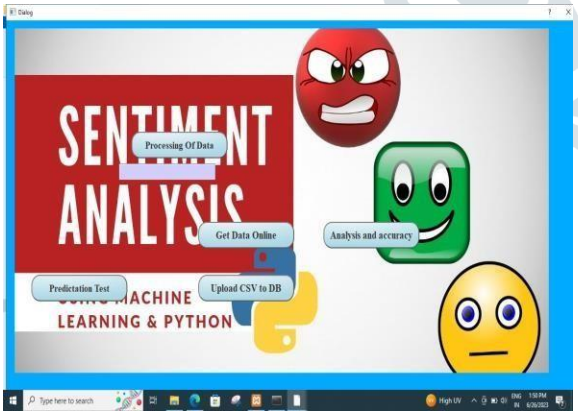

Figure 4: Main Interface

6.3.2 : Give comment for predictation test

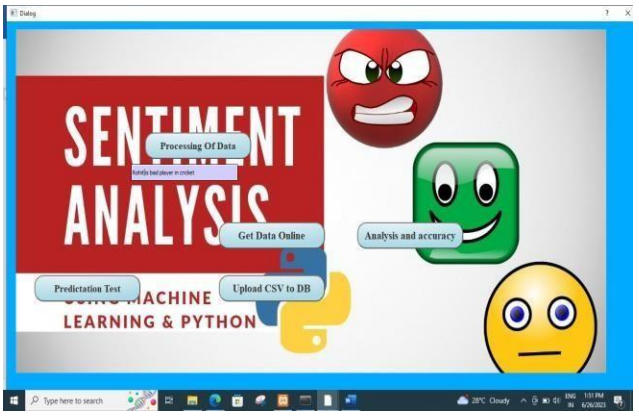

Figure5:Give comment

6.3.3 :Prediction on comment

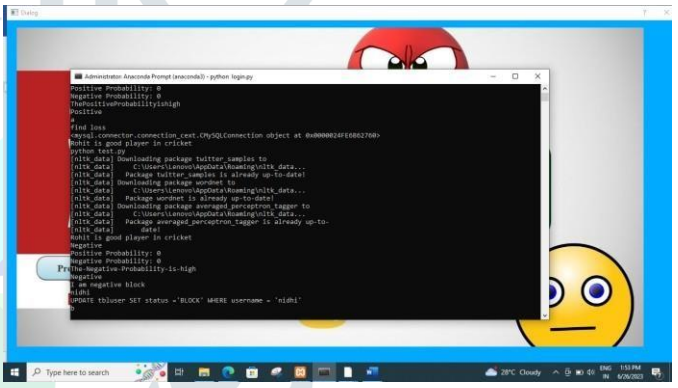

Figure 6:Prediction on comment

6.3.4 :Probability Prediction on comment

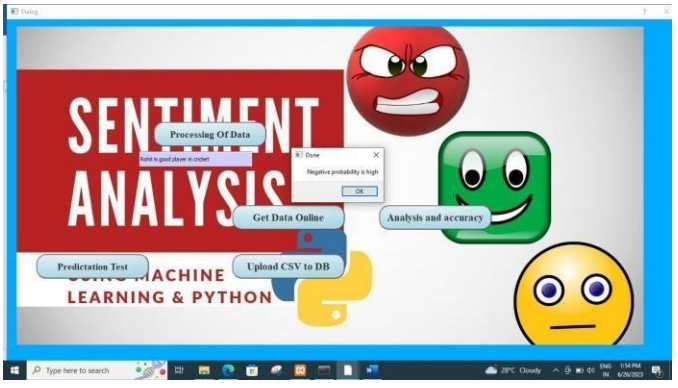

Figure 7:Probability Prediction

6.3.5 : Accessing Online Comments

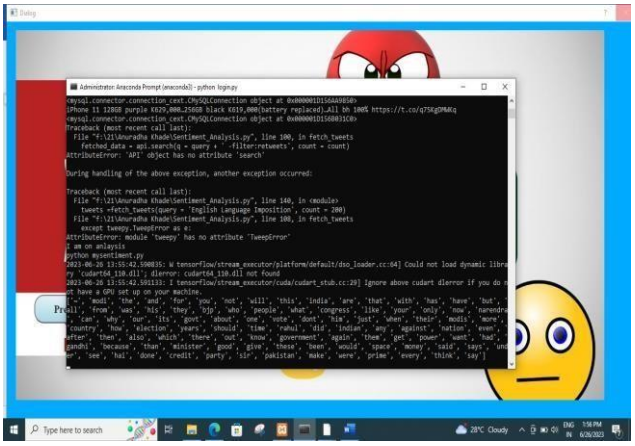

Figure 8:Access online comments

6.3.6 :Training and testing of precision recall and accuracy

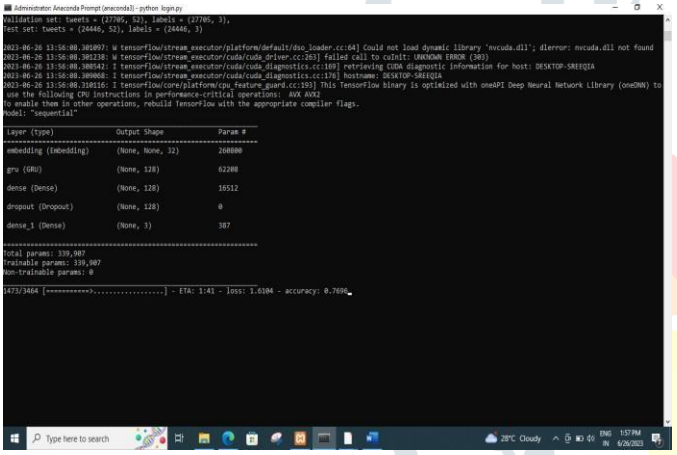

Figure 9: Training and testing of precision recall

6.3.7 :Graph of Accuracy

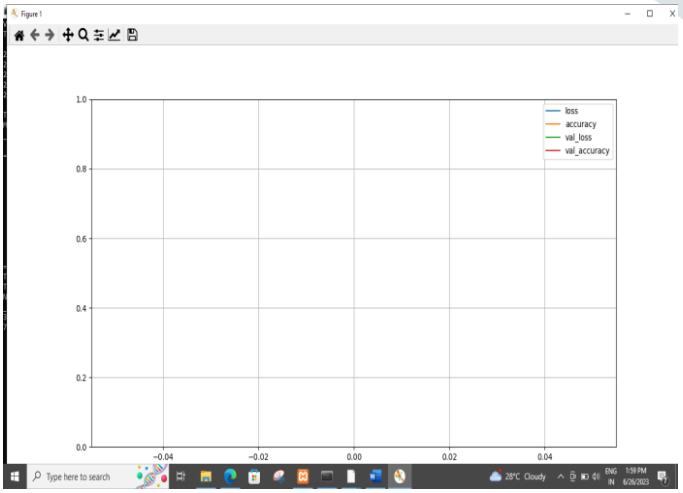

Figure 10: Graph Of Accuracy

6.3.8 :Prediction of Precision and Recall

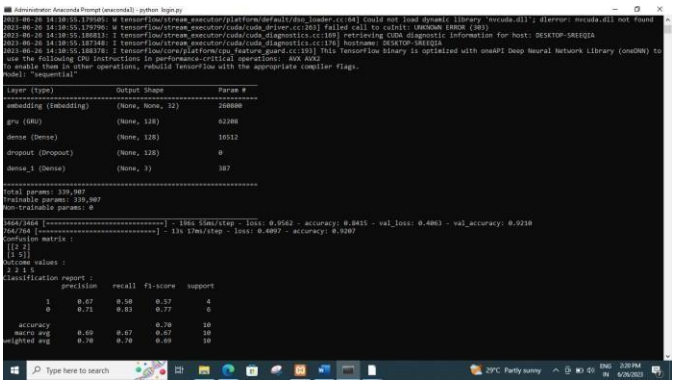

Figure 11:Prediction of Precision and Recall

6.3.9 :Graph Of Precision And Recall

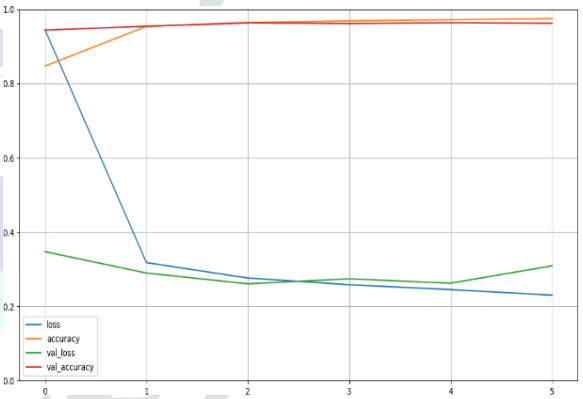

Figure 12: Graph Of Precision And Recall

VII. CONCLUSION:

Applying sentimental analysis to extract the sentiment became an important work for many organizations and even individuals. Sentiment analysis is an emerging field in decision making process and is developing fast. Our project goal is to analyze the sentiments on a topicwhich are extracted from the Twitter and determine its nature (positive/negative/neutral) of the defined topics. The development of techniques for the document-level sentiment analysisis one of the significant components of this area. Recently, people have started expressing their opinions on the Web that increased the need of analyzing the opinionated online contentfor various real-world applications. A lot of research is present in literature for detecting sentiment from the text. Still, there is a huge scope of improvement of these existing sentiment analysis models. Existing sentiment analysis models can be improved further with more semantic.

VIII. ACKNOWLEDGEMENT:

I would like to express my sincere thanks to Mrs. Hafsa Shoyeb Majgaonkar, Assistant Professor, Department of Computer Science and Engineering, Nanasaheb MahadikCollege Of Engineering, Peth. for her motivation, useful suggessions and guidance which truly helped me in improving the quality of this paper. Also would like to express my thanks to Principal Prof .Dr .B. Shrinivasa Varma, for his constant encouragement and support for carrying out this work.

REFERENCES:

[1] M. Rahbari, S. Rahlfs, E. Jortzik, I. Bogeski and
K. Becker, "H2O2 dynamics in the malaria parasite Plasmodium falciparum", PLoS ONE, vol. 12, no. 4, pp. 134-140, Mar. 2022.
[2] A. Bagchi Anjum, A. Parveen and R. Katarya, "Impact of Meteorological Parameters on COVID-19 Outbreak Using Machine Learning Techniques", 2021 12th International Conference on ComputingCommunication and Networking Technologies (ICCCNT), pp. 1-6, Jul. 2022.
[3] D. M. Morens, G. K. Folkers and A. S. Fauci, "What is a pandemic?", Journal of Infectious Diseases, vol. 200, no. 7, pp. 1018-1021, Oct. 2022.
[4] R. Feldman, "Techniques and applications for sentiment analysis: The main applications and challengesof one of the hottest research areas in computerscience", Communications of the ACM, vol. 56, no. 4, pp. 82-89, Apr. 2022.
[5] D. M. H. Thomson and C. Crocker, "A data- driven classification of feelings", Food Quality and Preference, vol. 27, no. 2, pp. 137-152, 2022.
[6] S. Chowdhury and M. P. Schoen, "ResearchPaper Classification using Supervised Machine Learning Techniques", 2022 Intermountain Engineering Technology and Computing IETC 2022, pp. 1-6, Oct. 2022.
[7] P. Kedia, Anjum and R. Katarya, "CoVNet-19: A Deep Learning model for the detection and analysis of COVID-19 patients", Applied Soft Computing, vol. 104, Jun. 2021.
[8] J. P. Ryans, "Textual classification of SEC comment letters", Review of Accounting Studies, vol. 26,no. 1, pp. 37-80, Mar. 2021.
[9] A. A. Soofi and A. Awan, "Classification Techniques in Machine Learning: Applications and Issues", Journal of Basic & Applied Sciences, vol. 13, pp.459-465, 2020.
[10] C. F. Chude and A. Q. Ezeoke, "PSYCHOLOGICAL EFFECT OF PANDEMIC COVID-19 ON FAMILIES OF HEALTH CARE PROFESSIONALS", British Journal of Psychology Research, vol. 8, no. 2, pp. 1-7, 2020.
[11] JurgenSchmidhuber, JyotikaPruthi, Deep learning in neural¨ networks: An overview. Neural networks, 61:85–117, 2020.
[12] SoroushVosoughi, Helen Zhou, and Deb Roy. Enhanced twitter sentiment classification usingcontextual information. arXiv preprint arXiv:1605.05195,2020
[13] A. Balahur,J. Hermida and A. Montoyo, 'Buildingand Exploitin Emotinet, a knowledge base for emotion detection based on the appraisal theory model', Affective Computing, IEEETransactions, vol. 3, 188101, 2020
[14] G.Vinodhini, Bhoomika Gupta and R. Chandrasekaran,'Sentiment analysis and opinion mining: A survey',International Journal, vol. 2, 6, 2020.
[15] Nancy Lazarus. 21 July 2019, “5 Key challenges ofsentiment analysis”, Blog.Ad weeks. Available
2020 12th International Conference on Computing Communication1, 6, 2020