1. INTRODUCTION

Web 3.0 is reshaping the social media landscape. In addition to using social media to engage, proportionate facts, and specific personal reviews, organizations also can use social media to communicate, understand, and improve their products and services. The use of Social media is increasing day by day.

Every day, it is anticipated that there will be up to 3.97 billion social media users on the planet by 2025. A variety of information is uploaded and shared across social media platforms in the form of text, videos, images, and audio. Social media is an enormous resource of raw and unprocessed data, and advances in technology, mostly in the field of machine learning and AI, are enabling these items to be processed and consolidated into valuable information from which organizations can gain the greatest competitive advantage.

In recent years, sentiment analysis has acquired sizable popularity among scholars,organizations, governments, and agencies (Sánchez-Rada and Iglesias 2019). The net's increasing reputation has extended it to the popularity of the primary supply of normal understanding. We need to use statistics generated by the user to assess it robotically as a way to hold a song of public interest and enhance decision-making. As a result, sentiment analysis has become more and more well-liked among research groups in recent years. Opinion evaluation or opinion mining are different phrases for sentiment analysis. The sentiment analysisproject has recently evolved in recognition.

The increasing use of social networking websites has given rise to a number of occupations focused on examining with the purpose to extract beneficial facts. Sentiment evaluation is the procedure for identifying emotions depicted by using text primarily based on the content. Sentiment analysis is a department of NLP, and given the lengthy and outstanding records of public opinionin selection-making, there must be numerous early publications on the issue. Nevertheless, it continues to broaden sentiment analysis into the next generation.

1.1 Problem Definition

Identifying if the expressed thought in a document, sentence, or entity feature or aspect is positive, negative, or neutral is a fundamental task in sentiment analysis.

People are using social media, including Twitter, which produces huge quantities of opinion texts in the form of tweets that are available for sentiment analysis, due to the rapid developmentof the internet. This interprets to a huge extent facts from a human point of view which makes itdifficult for extracting sentences, examine them, analyze tweets by way of tweets, explain them, and get them organized into a comprehensible layout in a well-timed way.

In this module, we have developed an algorithm based on sentimental analysis which will help us to know which text is positive or negative based on certain keywords present in a sentence. We will discuss the entire process in the upcoming pages.

1.1 Project Overview/Specifications

Sentiment analysis in Twitter is a discipline that has recently drawn a study hobby. One of the most well-known microblog sites where users can post their ideas and opinions is Twitter. Twitter's sentiment classification addresses the issue of live tweeting in light of the opinions theyexpress. This survey presents a top-level view of the subject with the aid of investigating and inbrief explaining the algorithms that have been expressed for sentiment evaluation on Twitter. The supplied research is labeled in line with the technique they observe. Further, we talk about areas related to sentiment analysis in Twitter which includes Twitter opinion retrieval, and monitoring sentiments over the years, The tasks of irony identification, emotion recognition, and tweet sentiment assessment have recently drawn much attention. Additionally, a brief list of sources used in the literature on Twitter sentiment analysis is given. The presentation of the suggested ways for sentiment evaluation in Twitter, their classification according to the methodology theyemploy, and the debate of latest studies tendencies of the subject and its associated fields.

1.2 Hardware Specification

RAM : At least 2GB ram(4GB ram recommended)Operating system : Windows 10/11

Processor: intel core i3 processor or aboveInternet speed: 5-10mbps

Cache: 512KB

1.3 Software Specification

- Django
- Dataset - twitter_sample
- Python 3.0 or above
- NLTK package
- Visual Studio Code
- Google Colab

2. LITERATURE SURVEY

In this section, we have discussed the existing system, proposed system, and feasibility study

2.1 Existing System

At various granularities, sentiment analysis is being treated as one of the tasks in natural language processing. It was originally a document-level categorization task, but more recently, it has been handled at the phrase level.

The three main models in sentiment analysis that are used to determine whether a sentence is positive or negative are SVM, Naive Bayes, and Language Models (N-Gram).

Support vector machine (SVM) could be a methodology for the classification of each linear and nonlinear knowledge. and also assert that SVM beats other classifiers. SVM is renowned for giving the best outcomes.

They explore parts-of-speech (POS) features in the feature space along with a Unigram, Bigram model. They note that the unigram model outperforms all other models. Particularly bigrams and POS traits.

In machine learning, the Naive Bayes classifiers are a family of simple probabilistic classifiers created by the application of Bayes' theorem with strong (naive) independence assumptions across the features. The number of variables (characteristics) in a learning problem is linearly inversely related to the number of parameters used for naive Bayes classifiers.

Moreover, they use search queries to acquire flawed data for testing and training. We insteadshow features that perform much better than different model baselines. Furthermore, we study a different method of data representation and discover that it performs massively better than the other models. The fact that we present findings on manually annotated data that is free of any known biases are another addition to this research. Unlike data gathered using specific queries, our data is a random sampling of live tweets. Due to the amount of our hand-labeled data, we can do cross-validation tests and examine how the classifier performs differently between folds.

The likelihood of Logistic regression may predict a result just with two values. (i.e. a dichotomy). Using one or more predictors, the prediction is based on data (numerical and categorical). For two reasons, estimating a binary variable's value using linear regression is inappropriate. Since there is only one of two possible values with each trial, the results in the dichotomous experiments will not be distributed regularly about the expected line.

2.2 Proposed System

In the projected system, looking out the data supported class and keywords from the info is performed. Looking out keywords is one of the toughest tasks due to the presence of multiple languages and also the bad words employed by the users. Within the projected system, the first step involves an assortment of information from totally different sources and creating it as a piece of information the associated information which was before the next step once everything is set. The (NaturalLanguage Processing) algorithm, which is used in the third stage, does sentiment analysis using numeric statistics. a certain emotion price mistreatment information processing is used as a weight consider sentiment analysis. Within the 4th step, the same knowledge is known and analyzed, and then deploying an internet application, the ultimate result, is that area unit suggestions are the area unit available for the problems occurring within such a method. The tweets area unit collected supported the mix of Keywords and classes provided by a user. Within the next step, all the info is removed from unnecessary words, symbols, and characters during pre-processing. In our proposedsystem pre-processing consists of three major steps which are as follows:-

- At first remove special characters like #,$,% and many others.
- use stopwords to ignore words like the, an, was and steamers to tokenize words intostrings.
- Converting the uppercase to lowercase was the next step.

In sentiment analysis, informatics analyses the mood of the accumulated information by acting by the subsequent steps:-

- In the first step it does tokenization.
- Then splitting the dataset into two parts one is a test set and the other one is a train set.
- Next step parses the sentence for syntactic analysis.
- The emotional value of the tweet supporting the outcomes of the preceding steps is then decided.

The final step is to style an online web application for providing the final output for the users and recommend a few other comments or results for the analyzed text.

- Get the positive data from the positive_tweets.json file to analyze the result.
- Add values comparator algorithm to the gathered positive data. to generate a list of suggestions provided by many users.

At first, the information ought to be collected from the dataset and a few sources. The set of data containing the collected information is pre-processed and parsed to eliminate common incorrect words, symbols, characters, and numbers, and to convert upper-case letters to lower-case letters.After pre-processing, the emotions are analyzed by an exploitation language process tool. Each sentence is given sentiment worth and supported by this sentiment worth the information is set as positive or negative. Each positive and negative information is analyzed and the same data is known.

2.3 Feasibility Study

A feasibility analysis is indeed a quick evaluation that looks into the data of potential users and ascertains the resources needed the costs, the advantages, and the viability of the suggested system.

Analyzing the issue and compiling all pertinent data about the project are both parts of the feasibility analysis process. The main goal of the feasibility analysis is to ascertain if or not In terms of economic viability, technical viability, operational viability, and schedule viability, the project would be feasible. It confirms sure the project's necessary input data are available.

There are different types of feasibility explained below:

2.3.1 Technical Feasibility

We must be extremely clear about the technologies that will be needed before we start the projectin order to construct the new system. Is the necessary technology readily available? Technically speaking, our "Sentiment Analysis" approach is feasible because all the necessary instruments are readily available. Python is simple to work with.

2.3.2 Operational Feasibility

The Project being proposed is solely advantageous if and only if it can be transformed into using information systems that would satiate operational needs. This feasibility test, put simply, inquires as to whether the system will function after development and installation. Do any significant obstacles to implementation? A streamlined application that analyzes a given text was proposed.It is easier to use and works on all Python platforms.

2.3.3 Economic feasibility

Economic feasibility is to balance the benefits of having the new system in place against the expenses of designing and implementing the new system. The senior management has an economic reason for the new system thanks to this feasibility study. A Simple economic analysisthat provides a direct comparison of costs and benefits is significantly more effective and more significant in this situation. Furthermore, this serves as an easy benchmark in order to match costsas a project moves along. The automation of certain processes may result in a variety of intangible advantages. These could lead to greater advancements in information timeliness, decision-making, and product quality. They could also speed up processes, increase operationalcorrectness, and provide better record-keeping and documentation. A precise set of findings areprovided by this application. It is inexpensive to create an application.

3. SYSTEM ANALYSIS AND DESIGN

In this section we have discussed the Requirement Specification, Design Test Steps Criteria, algorithms, and testing processes.

3.1 Requirement Specification

The system’s capacity to fulfill users’ demands is known as system demand. When conducting system requirement analysis, needs, requests, and wants are grouped. into purposeful needs and nonfunctional needs.

3.1.1 Requirement Definition

We have a better understanding of the requirements the existing system needs after a thorough investigation of the system's issues.

3.1.2 Functional Requirement

Functional requirements are attributes that any system must have in order to meet user expectations and fulfill business requirements.

The following are the functional specifications for the system must have:

- After retrieval, the system processes new tweets that were stored in a database.
- The gadget expects to be allowed to analyze statistics and classify each tweet's tone.

3.1.3 Non-Functional Requirement

The non-functional requirement includes a description of the system's attributes, features, and characteristics as well as any limitations that might be placed on the proposed system's capabilities. The main focus of the non-functional criteria is performance, information, economy, control, and security efficiency.

3.2 Design and Test Steps / Criteria

In the design method, the model is classified into 2 sets of data used as a training set and a testing dataset. Training data is used to train a classifier, an analysis set of data is used to evaluate the classification algorithm using 20% of the information from the training dataset, and a testing dataset is used to keep the classifier from automatically classifying the Twitter posts into one of four selected topics.

3.2.1 Development Phase and flowchart explanation: -

In this flow chart, the entire code flow is explained. Starting from importing libraries, downloading datasets removing stop words, and creating different sets for training and testing to building the model and applying logistic regression and getting the output as positive or negative.

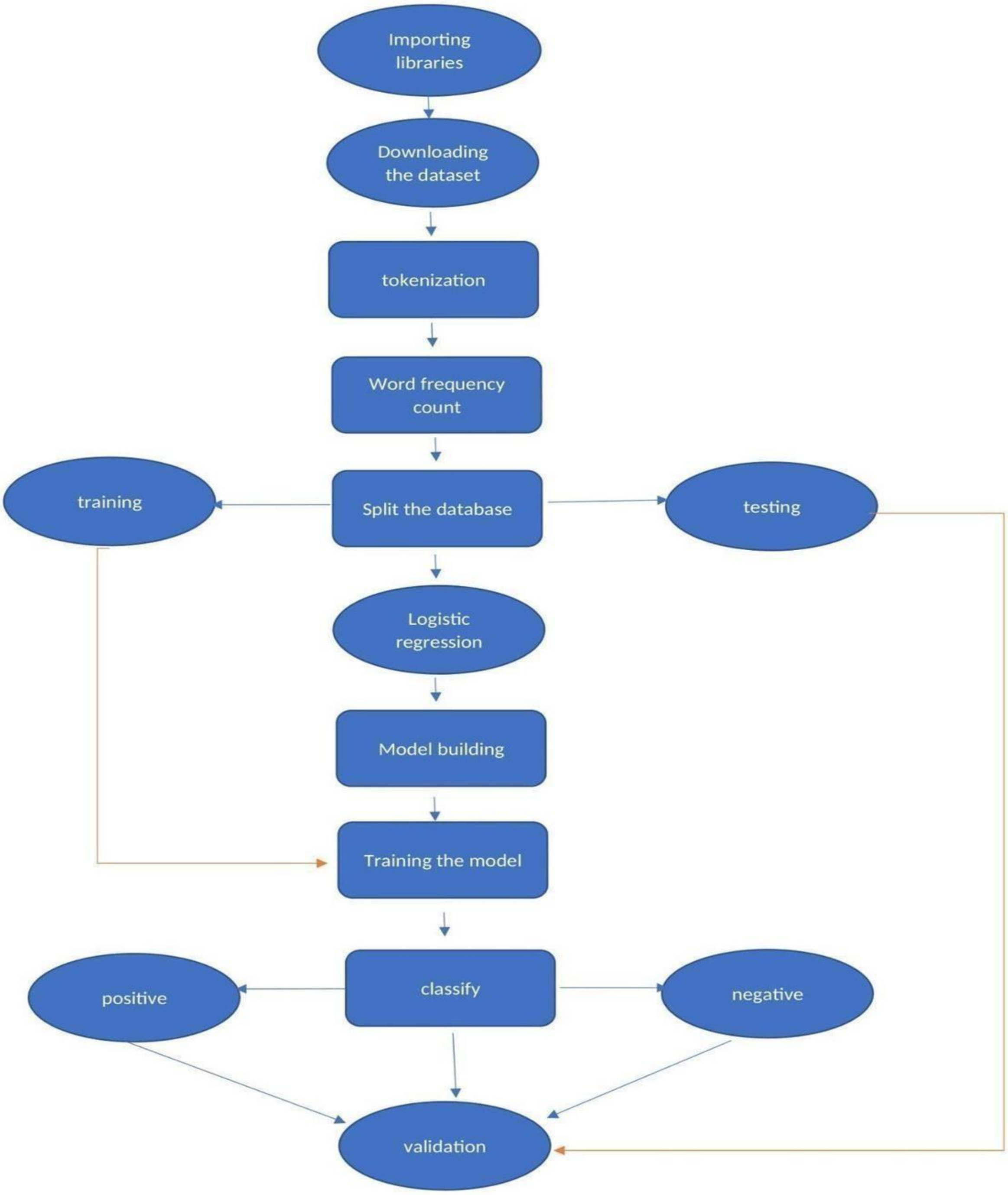


Figure 3.1 Flow Chart Diagram

1. **Fetching Tweet:-**The Twitter API is used, and additional filters depending on the topics'keywords are applied.

2. **Pre-processing:-This** his is taken into consideration as a critical segment as it prepares

the unstructured statistics for processing. If we have now no longer completed this step properly, then there may be an excessive opportunity that we may also address scattered and faulty information. So, the purpose of this segment is to do away with unrelated textual content from the Twitter posts like unique characters and punctuation which does now no longer upload any price to calculate sentiment.

- Due to privacy concerns, when assembling the tweets, any Twitter debts are hidden. Since the usernames of Twitter accounts (@user) aren't necessary for identifying the sentiment, we remove them entirely from the aggregated tweets.
- Punctuation, numerals, and other special characters are removed: In this phase, all text other than hashtags and characters is replaced with blank spaces.
- Short words should be eliminated in this step, thus we must be careful to select the length of each word. Thus, we decide to eliminate from the tweet any word length that is three characters or fewer. A couple of examples are the emotionlesswords "huh" and "oh."
- Tokenization: Tokens are single words, and tokenization is the process ofdividing a string into tokens. Let us take a tweet for example “He is good at singing”. After Tokenization, it looks like [‘He’, ‘is’, ‘good’, ‘at’, ‘singing’].
- Stemming is a technique for returning a word back to its underlying word by deleting the suffix. The normalizing approach known as stemming minimizes the amount of computations needed in natural language processing. For instance,the root word sing is shared by the words sing, singer, singing, and sings.

3. Text Feature Extraction:- Using the Bag of Words model, turn a batch of tweets into afeatures vector in real number form. The set of feature vectors is then passed to the Logistic Regression classifier. We have used the word frequency to count the number of positive and negative words and set it to a dictionary by the name freqs.

4. Train The Model (Use Logistic Regression):-

Logistic regression is known as the function used in the middle of the technique, the logistics evolved by statisticians to describe the homes of the populace increase in ecology, which is expanding swiftly and wearing out the environment.

Its S-shaped curve proves in figure 3.2 that Any real-valued amount can be convertedinto a price between zero and one, but never precisely at those ranges.

$$S(X) = 1 / (1 + e^{-x})$$

(1)

In (1), e is the Euler’s constant and value(x) is the number price that we want to convertand x is any real number.

This is a plot of the numbers between -5, and 5 converted into the variety 0 and 1 using thelogistic feature.

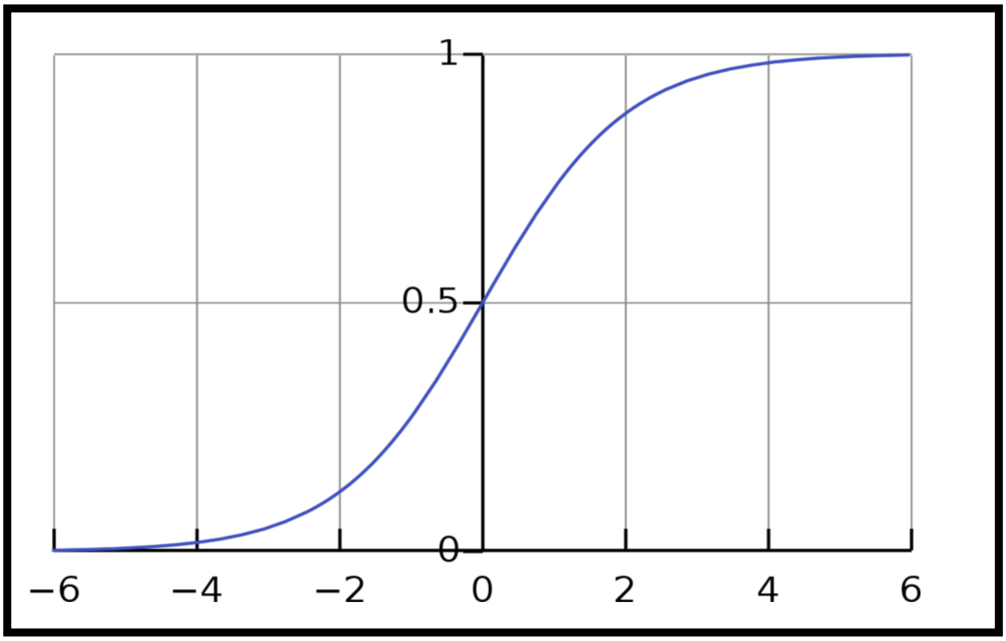


Figure 3.2 S curve

Similar to linear regression, logistic regression uses an equation as an example. To forecast an output value, enter data (x) are combined linearly with weights or coefficient values (referred toas the Greek capital letter Beta) (y). The fact that the output cost being modeled is a binary cost(0 or 1) rather than a value is a critical difference from regression analysis. Although logistic regression is a linear method, the predictions have changed how the logistic characteristic is used. As a result, unlike linear regression, we will no longer be able to understand the predictions as a linear aggregate of the inputs. Starting from before, the version might be phrasedas follows:

$$P(X) = e(b0 + b1*X) / (1 + e(b0 + b1*X))$$

(2)

In the above equation (2) wheree is the Euler constant

b0 and b1 are the coefficients

Cost function: -

In Linear regression, we got to understand the cost function J(θ), It stands for an optimization goal. In order to build an accurate model with minimal cost, we establish a cost function and minimizeit with the least amount of error.

m

$$J(\boldsymbol{\theta}) = 0.5 * \sum ((h_{\boldsymbol{\theta}} (x^{(i)}) - y^{(i)})^{2}$$

(3)i=1

in the above equation (3)j($\boldsymbol{\theta}$) is the cost function

$\boldsymbol{\theta}$ is the slope

H $\boldsymbol{\theta}$ is the output of the hypothesis

If we try to use the linear regression cost function in 'Logistic Regression,' it will be useless because it will end up being a Finding the global minimum and minimizing a non-convex function with numerous local minimums is highly challenging.

Gradient Descent:-

Gradient descent's main goal is to reduce the cost value. i.e., min J(θ).

m

$$\boldsymbol{\theta}j \leftarrow \boldsymbol{\theta}j - (\alpha/m)* \sum((h_{\boldsymbol{\theta}} (x^{(i)}) - y^{(i)}) * x_j^{(i)}$$

(4)i=1

in equation no 4, where

$\boldsymbol{\theta}$ is the slope

m is the total number of features

α is the learning rate

It has an associate degree Our goal is to succeed in the lowest of slopes, thus must imagine ourselves just at top of a steep natural depression, stuck, and wearing blinders. Everyone would feel the slope of the area of the earth around them. Taking a step is comparable to at least one iteration of the parameter update, and this action is comparable to clever gradient descent.

3.2.2 Design Modeling:-

This specific section can shortly justify the demonstration of all the processes and datasets for the tweet classification application modeling utilized in the submission. Once Twitter has properly downloaded the tweets, victimization Twitter API, these data can go to the preprocessing method. After the preprocessing method is accomplished, the classifier that uses supply can be trained using tweets from the training dataset using the training method. Regression. In order todetermine the trained classifier's accuracy rate, the sensitivity confusion matrix is used to evaluate the trained classifier.

1.1 Algorithms and Pseudo Code

Here the algorithm and pseudo code is explained step-by-step

Step 1. importing the necessary libraries like nltk, re, and stop words

Step 2. filter out all the unwanted words that are not useful

Create a function filter_words

remove the stop words and remove the punctuation

Step 3. tokenizing helps us to get individual strings known as word
tokens using tokenizer. tokenize to tokenize the tweet into word tokens

Step 4. **empty array initialized**

words_clean.= []

Step 5. condition to check if the word is not in stopwords or in punctuation

If the Word not in string and not in English punctuation find the root word and store it in the words_clean array

Step 6 .definining a new method pair_freqs

def pair_freqs

Step 7. filter out words and increase the positive count by 1 and set the negative count to 1

for any word in the zip

for word in filter_words(word):

set the value of the pair as (word,y)if the pair is in freqs:

increase the count freqs[pair] by 1

else:

set the count freqs[pair] as 1

return freqs

Step 8. set two empty arrays to the JSON files respectively and Set their values from positive_tweets.json and negative_tweets.json respectively. positive_words = twitter_samples.strings('positive_tweets.json')

negative_words = twitter_samples. strings('negative_tweets.json')

Step 9. Initialize them with values ranging from 0 to 4000 and 4000 till the end.

Step 9. Create 4 sets. train_y,train_x,test_y,test_xand do the following

train_x= train_positive + train_negative

test_x= test_positive + test_negative

Step 10 Create a new function sigmoid(z)neg = np.negative(z)

return sig_x = 1 / (1 + np.exp(neg))

Step 11. gradient descent formula to get the optimal cost

Create a new function gradient_Descent(x, y,a,b,c)
for i in range 0 to c
initialise cost = -1. / m * (np.dot(y.transpose(), np.log(sig_x)) +np.dot((1 -y).transpose(), np.log(1 - sig_x)))
a = a - (b / m) * np.dot(x.transpose(),(sig_x - y))
return cost

Step 12. getting the values of freqs dictionary and set it to the array
Create a new function extract_features
initialize array of size 1,3 and set it as x[0, 0] = 1for word in filtered_words(tweets)

x[0, 1] = x[0, 1] + freqs.get((word, 1.0), 0)

x[0, 2] = x[0, 2] + freqs.get((word, 0.0), 0)return x

Step13.Create a new function predictVal = predict_word()

Step13.Create a new function predict Val = predict_word()
if val is greater than 0.5 return positive
if val is less than 0.5 return negative
return neutral

3.3 Testing Process

Now comes the most important part of the project i.e testing and ensuring if it works or breaks

Basically run the project, input the tweet in the text box, and hit on the analyze button to get the output.

1. The first step would be to calculate the prediction generated by the sigmoid function h(x).
2. set a threshold value (0.5) and check the prediction value.

a. if the predicted value is more than the threshold value then it is a positive tweet

b. else it is a negative tweet

3. Finally, compute the accuracy of the model over the validation set.

4. accuracy is the no of times the actual value matches the predicted value and canbe calculated

   using the below formula

$$\text{accuracy} \rightarrow \sum (\text{pred}^{(i)} == \text{yval}^{(i)}) / m$$

(5)

i=1

in the (5):

pred(i) is the predicted value at ith intervalYval is the actual value at ith interval

Accuracy is the total outcome / number of operations

4. RESULTS / OUTPUTS

Here getting the input from the user and showing the output accordingly by taking into consideration two test cases.1st test case will be of positive nature and the 2nd test case willbe of negative nature.

4.2.1 Test case

When the input is completely positive data, i.e. when the data collected about the productor anything is completely positive. The sentiment analysis system's output is as follows:

Input:- System asks the user to input the data in the textbox and clicks on analyzebutton where the input text: is “i am going to start reading the Harry Potter seriesbecause that is a good story.”

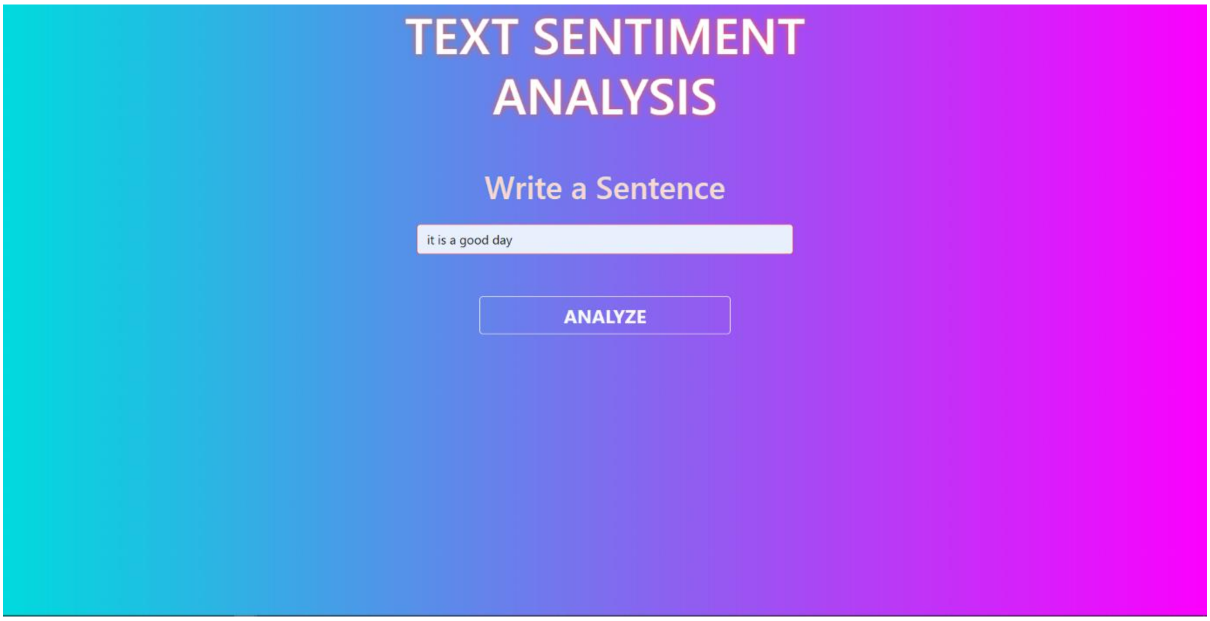


Figure 4.1 Text Entering

Output:-

The algorithm detects the nature of the text and shows output as positive.

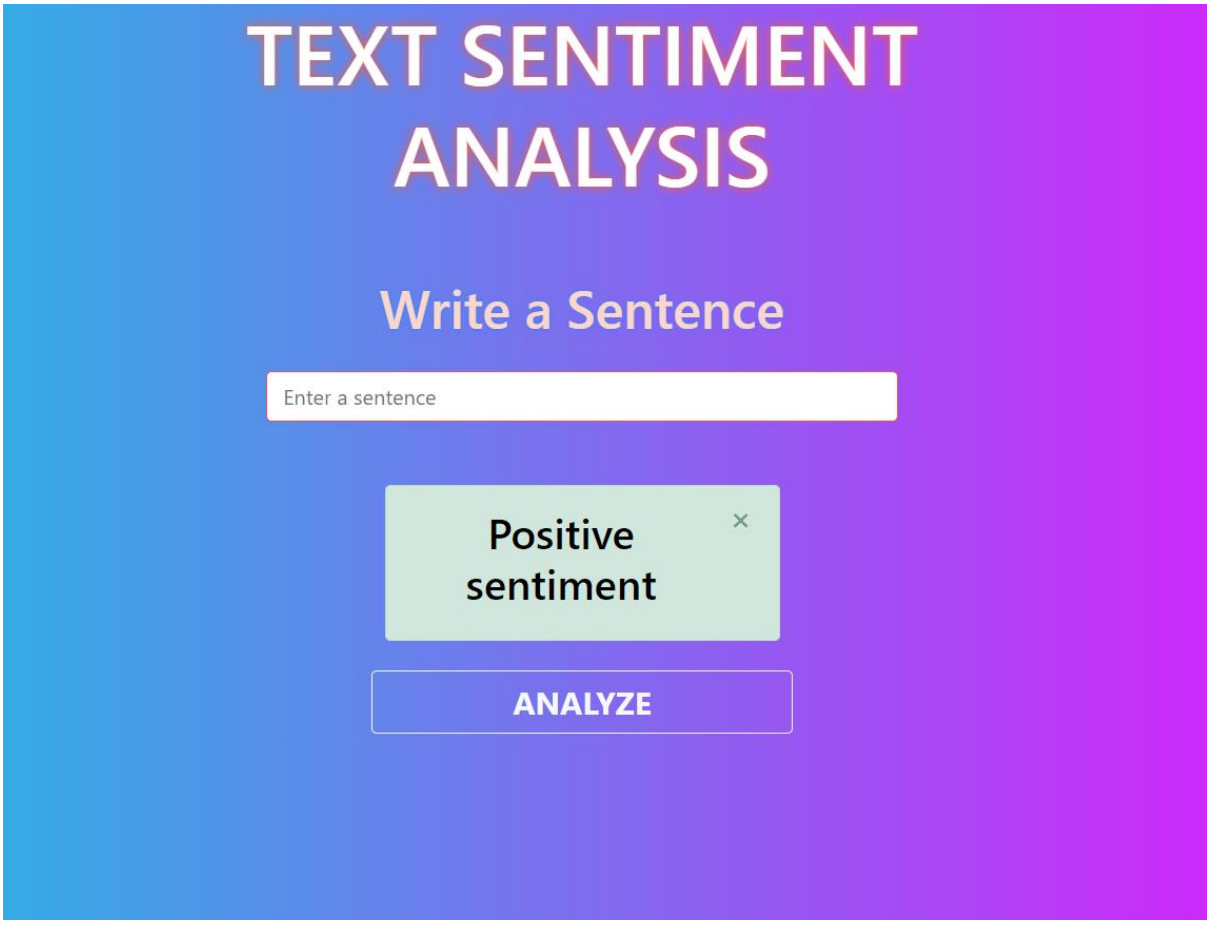


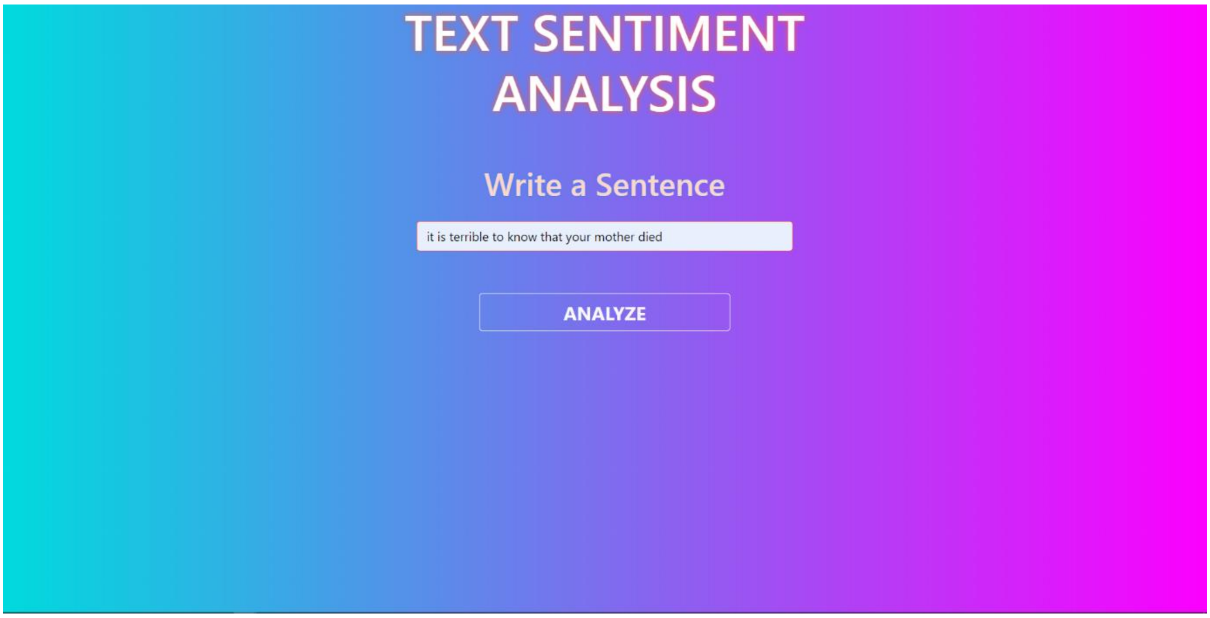
Figure 4.2 Result as positive

4.2.1 Test case 2

When the input is completely negative data, i.e., when the data collected about the product oranything completely negative. The sentiment analysis system's output is as follows:

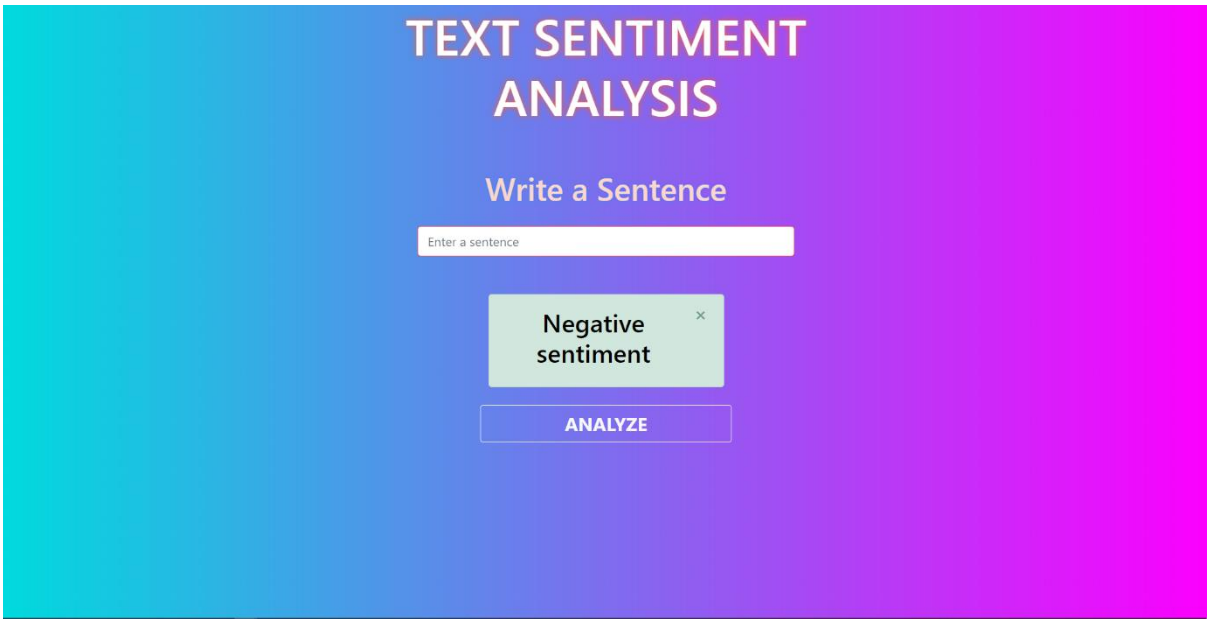
Input:-

System asks the user to input the data in the textbox and clicks on analyze button where The input text is "**for those who are Harry Potter ignorant, the villains of this movieare awful creatures called dementors**"



it is terrible to know that your mother died

Output:-The algorithm detects the nature of the text and shows output as negative



Negative sentiment

Figure 4.4 Negative as result

5. CONCLUSIONS AND FUTURE SCOPES

5.1 Conclusions

So, from the above pages, it is known what are methodologies used in developing the project.To sum it up, logistic regression based model is used along with Django backend with twitter_dataset to develop this application. fetch data from the dataset and compared it with thetokenized strings that are obtained from the input string, preprocess it, remove unnecessary stopwords and then apply logistic regression on it to get the actual tone of the sentence be it positive or negative.

The only downside of the algorithm is it cannot detect neutral inputs. so try to plot the same usingthe sigmoid function and its value ranges from 0 to 1. A line is drawn in the middle of the graph to find the 0.5 points. It is concluded that whatever the model gives a value that lies on this point is ofneutral nature but it is difficult to predict which is actually positive or actually negative. Hence,won't be able to find the Neutrality of a sentence in the case of this algorithm.

5.2 Future Scopes

Talking about the future scopes, this Sentiment analysis will include and appreciate the significance of social media conversations, going away from the idea of the number of likes, shares, and comments on a piece. It can be set up as a third-party plugin that will help to filterout spam/hate/unnecessary content harmful to the audience. It is also planned to make this open source so that any developer can make any changes to the source code and make a better versionout of it. It can also have emoticons to express the tone of the tweet which will be more interactive and fun as compared to pop up texts. starting an awareness program where one can teach non-technical people how to use this and beware of mishaps in social media.

6. REFERENCES

[1]. Tyagi, A., & Sharma, N. (2018). Sentiment analysis using logistic regression andeffective word score heuristic. *International Journal of Engineering and Technology (UAE)*, *7*(2.24), 20-23.

[2]. Stine, R. A. (2019). Sentiment analysis. *Annual review of statistics and its application*,*6*, 287-308.

[3]. S Sindhura, S Phani Praveen, M.Aruna Safali, Nidamanuru Srinivasa Rao, "Sentiment Analysis for Product Reviews Based on Weakly-Supervised Deep Embedding", 2021 Third InternationalConference on Inventive Research in Computing Applications (ICIRCA), pp.999-1004, 2021

[4]. Ansh Gupta, Aryan Rastogi, Avita Katal, "A Comparative Study of Amazon ProductReviews Using Sentiment Analysis", 2021 International Conference on Advances in Computing, Communication, and Control (ICAC3), pp.1-6, 2021.

[5]. Marius Ngaboyamahina, Sun Yi, "The Impact of Sentiment Analysis on Social Mediato Assess Customer Satisfaction: Case of Rwanda", 2019 IEEE 4th International Conference on Big Data Analytics (ICBDA), pp.356-359, 2019.

[6]. Adwan, O., Al-Tawil, M., Huneiti, A., Shahin, R., Zayed, A. A., & Al-Dibsi, R. (2020). Twitter sentiment analysis approaches: A survey. International Journal ofEmerging Technologies in Learning (iJET), 15(15), 79-93.

[7]. Giachanou, A., & Crestani, F. (2016). Like it or not: A survey of twitter sentimentanalysis methods. ACM Computing Surveys (CSUR), 49(2), 1-41.