SENTIMENTAL ANALYSIS

Project Report

SUBMITTED IN PARTIAL FULFILLMENT REQUIREMENT FOR THE AWARD OF DEGREE OF
Bachelor of Technology

(COMPUTER SCIENCE & ENGINEERING)

SUBMITTED BY

Mubashir Buhari
Karthik Raja Nichenametla
Godavarthi Sai Nikhil

(UNIVERSITY ROLL NO.

210407
210371
210214)

UNDER THE SUPERVISION OF
Dr. NISHTHA PHUTELA
SCHOOL OF ENGINEERING AND TECHNOLOGY

BML MUNJAL UNIVERSITY™
FROM HERE TO THE WORLD

BML MUNJAL UNIVERSITY
Gurugram, Haryana - 122413
Dec 2022

CANDIDATE'S DECLARATION

I hereby certify that I have under gone six months industrial training at BML MUNJAL UNIVERSITY and worked on project entitled, "**SENTIMENTAL ANALYSIS**", ", in partial fulfillment of requirements for the award of Degree of **Bachelor of Technology** in name of the department at **BML MUNJAL UNIVERSITY**, having University Roll No.210214, is an authentic record of my own work carried out during a period from July,2021 to December, 2021 under the supervision of **Dr. NISHTHA PHUTELA**.

Godavarthi Sai Nikhil

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

(Dr. NISHTHA PHUTELA)

Person from Organization with designation

ABSTRACT

Sentiment analysis and opinion mining are two terms for the computer analysis of individuals' opinions, sentiments, attitudes, and emotions as they are expressed in written language. One of the most active research areas in text mining and natural language processing in recent years is this one. Its reasons for attractiveness are mostly influenced by two elements.

- First, it has many uses since views are fundamental to practically all human endeavors. Our behavior is significantly influenced by our activities. Every time we have to make a choice, we want to find out what others think.

- Second, it provides several difficult research issues that have never before tried prior to the year 2000. Opinions had a role in the absence of prior research. tried prior to the year 2000. Prior until now, there wasn't much opinionated text available in digital forms, which contributed to the lack of research.

So, it shouldn't be a surprise that the area's rapid growth and emergence coincide with those of online social media. Study on this subject has actually expanded beyond computer science to include management sciences and social sciences due of its importance for business and society at large. Prior to describing some recent work on modeling comments, discussions, and arguments, which is another sort of sentiment analysis and opinion analysis, I will first review mainstream sentiment analysis research.

ACKNOWLEDGEMENT

I am highly grateful to **Dr. NISHTHA PHUTELA,** BML Munjal University, Gurugram, for providing supervision to carry out the seminar/case study from July-December 2022.

(Dr. NISHTHA PHUTELA) has provided has provided great help in carrying out my work and is acknowledged with reverential thanks. Without the wise counsel and able guidance, it would have been impossible to complete the training in this manner.

I would like to express thanks profusely to thank **Dr. NISHTHA PHUTELA**, for stimulating me time to time. I would also like to thank entire team of BML Munjal University. I would also thank my friends who devoted their valuable time and helped me in all possible waystowards successful completion.

(GODAVARTHI SAI NIKHIL)

LIST OF FIGURES

TABLE OF CONTENTS

Chapter 1

1.1 PROBLEM STATEMENT

A Basic Task in sentimental analysis is classifying the tone of a text in a document sentence or future aspect level--whether the given text in a document or sentence is positive or negative. Understanding about Customer attitude and trends. Today, if any one wants to purchase a product or to give vote or to watch a movie, etc. then that person will first want to know what are other people reviews, reactions and opinions about that product or candidate or movie on social media websites like Twitter, Facebook, Tumbler, etc. So, there is a need of system that can automatically generate sentiment analysis from this huge amount of data.

1.2 OBJECTIVES

- Identifying the text's tone, emotion, attitude, and opinion is the goal of this project.
- In despite the challenges, we try to use a sentiment analysis model in this model that helps identify the sentiments of the Amazon reviews. considering the utilization of natural language processing.
- By utilizing LSTM and TensorFlow to create a deep learning model. also creating a website where users may gauge the mood of a sentence.

Chapter 2

Introduction to Organization:

BML Munjal University wants to revolutionize higher education in India by developing a top-notch and innovative environment for teaching, learning, and research. The Hero Group founded the BML Munjal University, which offers undergraduate and graduate programs. Imperial College London serves as its mentor.

Chapter 3

Introduction

In this chapter we are going to give the introductions on Sentiment Analysis, Python and Natural Language Toolkit (NLTK). Then we are explaining the objective of our thesis. After this we will discuss why there is a need of sentiment analysis and some of the applications of Sentiment Analysis which are used in our daily life. 1.1 Introduction to Sentiment Analysis Sentiment Analysis is process of collecting and analyzing data based upon the person feelings, reviews and thoughts. Sentimental analysis often called as opinion mining as it mines the important feature from people opinions. Sentimental Analysis is done by using various machine learning techniques, statistical models and Natural Language Processing (NLP) for the extraction of feature from a large data. Sentiment Analysis can be done at document, phrase and sentence level. In document level, summary of the entire document is taken first and then it is analyzed whether the sentiment is positive, negative or neutral. In phrase level, analysis of phrases in a sentence is taken in account to check the polarity. In Sentence level, each sentence is classified in a particular class to provide the sentiment. Sentimental Analysis has various applications. It is used to generate opinions for people of social media by analyzing their feelings or thoughts which they provide in form of text. Sentiment Analysis is domain centered, i.e., results of one domain cannot be applied to another domain. Sentimental Analysis is used in many real-life scenarios, to get reviews about any product or movies, to get the financial report of any company, for predictions or marketing. Twitter is a micro blogging platform where anyone can read or write short form of message which is called tweets. The amount of data accumulated on twitter is very huge. This data is unstructured and written in natural language. Twitter Sentimental Analysis is the process of accessing tweets for a particular topic and predicts the sentiment of the given text.

Chapter 3

Introduction to Project

3.1 Introduction to Python:

language which is used for this thesis. Python3.4 version was used as it is a mature, versatile and robust programming language. It is an interpreted language which makes the testing and debugging extremely quickly as there is no compilation step. There are extensive open-source libraries available for this version of python and a large community of users. Python is simple yet powerful, interpreted and dynamic programming language, which is well known for its functionality of processing natural language data, i.e., spoken English using NLTK. Other high level programming languages such as 'R' and 'MATLAB' were considered because they have many benefits such as ease of use but they do not offer the same flexibility and freedom that Python can deliver.

3.2 Introduction to NLTK:

Natural Language Toolkit (NLTK) is library in Python, which provides a base for building programs and classification of data. NLTK is a collection of resources for Python that can be used for text processing, classification, tagging and tokenization. This toolbox plays a key role in transforming the text data in the tweets into a format that can be used to extract sentiment from them. NLTK provides various functions which are used in pre-processing of data so that data available from twitter become fit for mining and extracting features. NLTK support various machine learning algorithms which are used for training classifier and to calculate the accuracy of different classifier. In our thesis we use Python as our base programming language which is used for writing code snippets. NLTK is a library of Python which plays a very important role in converting natural language text to a sentiment either positive or negative. NLTK also provides different sets of data which are used for training classifiers. These datasets are structured and stored in library of NLTK, which can be accessed easily with the help of Python.

3.3 Introduction to Flask

Here you will see about the Flask documentation. Start with Installation, then proceed to the QuickStart for a summary. There is also a longer Tutorial that demonstrates how to create a simple yet complete application using Flask. Typical patterns are described in the Patterns for Flask section. Each Flask component is covered in great length in the rest of the documentation, which also includes an API section with a comprehensive reference. The Werkzeug WSGI framework and the Jinja template engine are both necessary for Flask to function.

3.4 Introduction to SKLearn:

A free machine learning library for Python, Scikit-learn was formerly known as Scikits. learn and is also known as Supervised learning. Among the

Chapter 3

clustering, regression, and classification algorithms it provides are support-vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy. Scikit-learn is a NumFOCUS fiscally sponsored project.

Sentimental analysis equation:

$$h_t = \sigma(W^H h_{t-1} + W^X x_t)$$

W = Represent weight matrices

We can conclude that ht is mostly influenced by ht-1 and unaffected by xt when the magnitudes of WH and WX are big and small, respectively.

GOAL OF THE PROJECT:

The most effective high-level programming language for NLP is Python. In order to process data in natural language, Python makes use of one of its packages called Natural Language Toolkit. The large number of corpora that NLTK offers makes it possible to train classifiers and perform all NLP operations on given datasets, such as tokenization, part-of-speech tagging, stemming, lemmatization, parsing, and sentiment analysis. Large datasets can be challenging to work with, but with NLTK, we can quickly classify our data and deliver more accurate results depending on several classifiers. Twitter is used to gather public opinions of these parties, which are then supervised machine learning classifiers categorize as positive or negative sentiments. These findings will inform us of what people are saying about these political parties. In order to accomplish this, a module that can conduct live sentimental analysis was developed. Using live graphs that show two sentiment categories, users can learn the trend of any live trending topic (positive and negative). Various machine learning classifiers can be used to further evaluate the module's accuracy and dependability. For many businesses and organizations, client reviews of a good or service are incredibly valuable sources of information. With the insights gathered from such an analysis, a company can pinpoint issues with its goods, anticipate trends before its rivals can, enhance communications with its target market, and learn valuable information about the effectiveness of its marketing initiatives.

Chapter 4

Literature Review

Many research has been done on the subject of sentiment analysis in past. Latest research in this area is to perform sentiment analysis on data generated by user from many social networking websites like Facebook, Twitter, Amazon, etc. Mostly research on sentiment analysis depend on machine learning algorithms, whose main focus is to find whether given text is in favor or against and to identify polarity of text. In this chapter we will provide insight of some of the research work which helps us to understand the topic deep.

Described a model that gathers tweets from social media platforms to give a view of business intelligence. The sentiment analysis tool has two layers in our framework, a layer for data processing and another for sentiment analysis. Data mining and data collection are dealt with in the data processing layer.

Explained a well-known microblog by the name of Twitter and developed models to categorize the "tweets" into positive and negative sentiment or neutral sentiment. The author develops novel models for two classification tasks: the first is a binary task that divides user sentiment into positive and negative classes, and the second is a three-way task that divides user sentiment into positive, negative, and neutral categories. The author uses two different kinds of models:

Using a tree-based model with 30 kernels, the unigram model.

Pak and Alexander and others suggested:

Author creates a sentiment classifier using the corpus that can identify the positive, neutral, and negative feelings present throughout the entire document. Experimental findings demonstrate that the suggested strategies outperform and are more effective than those that had previously been put forth.

P. Pang, L. Lee

They developed sentiment analysis as the pioneers. Their primary goal was to categorize material by overall attitude rather of just by topic, for example, by categorizing movie reviews as positive or negative. They use a machine learning algorithm on a database of movie reviews, and the findings show that these algorithms outperform those created by humans. They employ Naive-Bayes, maximum entropy, and support vector machines as their machine learning methods. They also get the conclusion that it is quite difficult to categorize sentiment after looking at a variety of elements. They demonstrate that the foundation of sentiment analysis is supervised machine learning methods.

Chapter 4

E. Loper, S. Bird et al

Natural Language Toolkit (NLTK) is a library which consists of many program modules, large set of structured files, various tutorials, problem sets, many statistics functions, ready-to-use machine learning classifiers, computational linguistics courseware, etc. The main purpose of NLTK is to carry out natural language processing, i.e., to perform analysis on human language data. NLTK provides corpora which are used for training classifiers. Developers create new components and replace them with existing component, more structured programs are created and more sophisticated results are given by dataset.

H. Wang, D. Can, F. Bar, S. Narayana et al

For the U.S. presidential elections in 2012, it was these scholars that suggested a mechanism for real-time analysis of public reactions. They gather the answers from the microblogging website Twitter. Twitter is one of the social media platforms where users may express their views, ideas, and opinions on any hot issue. Twitter comments from American election candidates generated a lot of data that was used to gauge public opinion of each contender and forecast who would win. The reactions individuals post on Twitter and the entire election cycle are connected in terms of feelings. They investigate sentiment analysis' impact on these public events as well.

O. Almatrafi, S. Parack, B. Chavan et al

They are the scientists who suggested a location-based system. They contend that Natural Language Processing (NLP) and machine learning algorithms do sentiment analysis. To extract an emotion from a text unit that comes from a machine learning specific place They research several uses for location-based sentiment. analysis employing a data source that allows for the extraction of data from several sources easily. A person may readily access the tweet location field on Twitter. script, meaning information (tweets) from a certain region may be gathered for identification patterns and trends.

Chapter 5

Methodology

Importing libraries and the Data Collection:

We have mainly used the library called "Tensor Flow". As per Project requirements dataset is required in a proper format for analyzing and preprocessing the model, So we have taken amazon review dataset using the library called "Tensor Flow".

Data Preprocessing:

Tokenization is the procedure of converting text into tokens before it is turned into vectors. It is also simpler to remove unnecessary tokens from the filter.
We have split our dataset into 2 parts i.e., Training Data, Test Data.
We will remove stop words from all the amazon reviews, now stop words are basically the most commonly occurring words in any language.
For Example: 'who', everything', 'Cutest', 'somewhat', 'Whatever', etc.

Detecting the sentiment:

Sentiment analysis is done at different levels using common computational techniques like Unigrams, lemmas, LSTM and RNN, and so on.

Classification of sentiment:

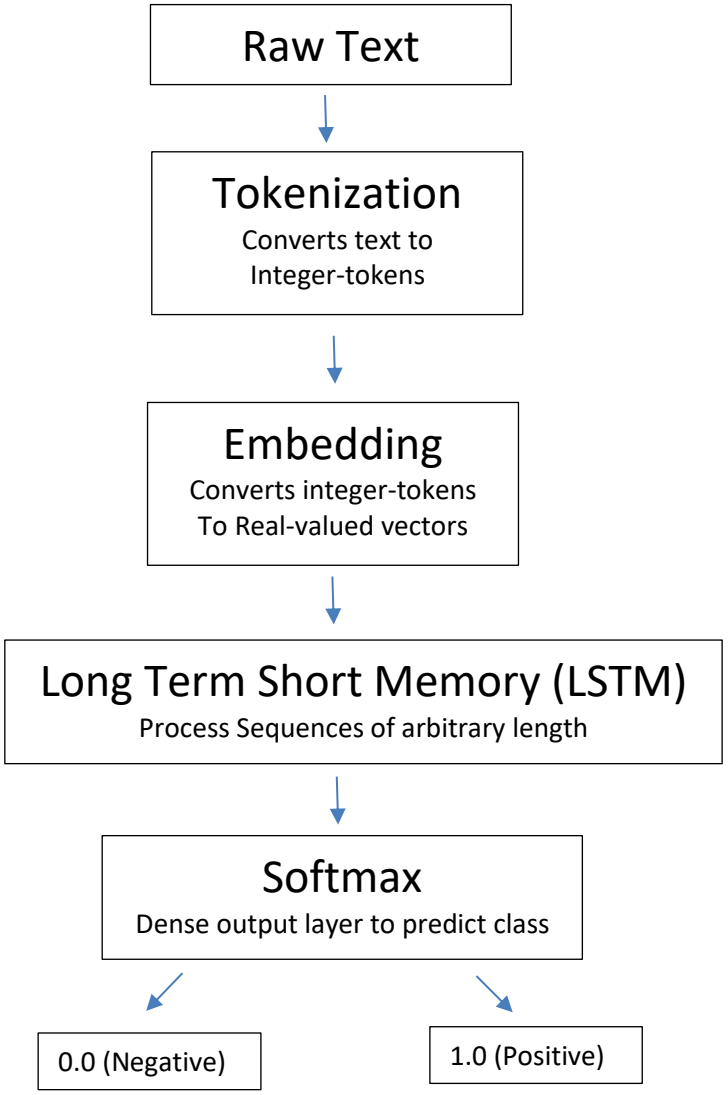
Sentiments can be broadly classified into two groups, positive and negative. At this stage of sentiment analysis methodology, each sentence detected is classified into groups-positive, negative.

5.1 Introduction to Languages (Front End and Back End)

For the Front-End part, we have chosen the HTML and CSS.
For the Back-End part, we have chosen the Flask and python.

Chapter 5

5.2 ML algorithm discussion



5.21 Raw Text:

We are using amazon mobile electronics reviews dataset which we acquired using TensorFlow used to train and validate our models. The dataset has individual product rating. Later we labelled them using our model with the help of ratings like >3 positive and <3 as negative. we convert the text to lower case and remove punctuation. Now we have to separate the individual reviews and store them in individual list elements. Like, [review_1, review_2, review_n]

Chapter 5

```
tfds.core.DatasetInfo(
    name='amazon_us_reviews',
    full_name='amazon_us_reviews/Mobile_Electronics_v1_00/0.1.0',
    description="""
    Amazon Customer Reviews (a.k.a. Product Reviews) is one of Amazons iconic products. In a period of over two decades since the first review in 1995, millions of Amazon customers
have contributed over a hundred million reviews to express opinions and describe their experiences regarding products on the Amazon.com website. This makes Amazon Customer Reviews a
rich source of information for academic researchers in the fields of Natural Language Processing (NLP), Information Retrieval (IR), and Machine Learning (ML), amongst others.
Accordingly, we are releasing this data to further research in multiple disciplines related to understanding customer product experiences. Specifically, this dataset was constructed
to represent a sample of customer evaluations and opinions, variation in the perception of a product across geographical regions, and promotional intent or bias in reviews.

    Over 130+ million customer reviews are available to researchers as part of this release. The data is available in TSV files in the amazon-reviews-pds S3 bucket in AWS US East
Region. Each line in the data files corresponds to an individual review (tab delimited, with no quote and escape characters).

    Each Dataset contains the following columns :
      marketplace       - 2 letter country code of the marketplace where the review was written.
      customer_id       - Random identifier that can be used to aggregate reviews written by a single author.
      review_id         - The unique ID of the review.
      product_id        - The unique Product ID the review pertains to. In the multilingual dataset the reviews
                          for the same product in different countries can be grouped by the same product_id.
      product_parent    - Random identifier that can be used to aggregate reviews for the same product.
      product_title     - Title of the product.
      product_category  - Broad product category that can be used to group reviews
                          (also used to group the dataset into coherent parts).
      star_rating       - The 1-5 star rating of the review.
      helpful_votes     - Number of helpful votes.
      total_votes       - Number of total votes the review received.
      vine              - Review was written as part of the Vine program.
      verified_purchase - The review is on a verified purchase.
      review_headline   - The title of the review.
      review_body       - The review text.
      review_date       - The date the review was written.
    """,
    config_description="""
    A dataset consisting of reviews of Amazon Mobile_Electronics_v1_00 products in US marketplace. Each product has its own version as specified with it.
    """
```

Fig. 5.21 Image of our Dataset.

5.22 Tokenizer:

The process of tokenizing, or dividing a string of text into a list of tokens, is known as tokenization. A word may be a token in a very specific situation if one were to think of tokens as pieces of that word.

All or any NLP jobs require tokenizing, which is the process of breaking a string into its intended component components. Because of tokenization, there is no singular right. The right algorithm will vary depending on the appliance. Since sentiment information is frequently sparingly and unusually represented, I suspect that tokenization is even more crucial in sentiment analysis than it is in other NLP applications. A single cluster of punctuation, such as >:-(, can be all that is needed to convey the full meaning.

5.23 Embedding:

Word embedding emerged from Natural Language Processing (NLP), a discipline that blends computational linguistics, artificial intelligence, machine learning, and computer science. The associations between words in textual data can be found using a text mining method called word embedding. The syntactic and semantic meaning of a word depends on the context in which it is used. The distributional hypothesis proposes that words have semantically similar meanings when they occur in comparable settings. Count-based embeddings and prediction-based embeddings are the two primary techniques for word embedding. Embeddings are used to capture language relationships. Embeddings are dense vector representations of the characters. It displays words in a highly organized way where closely related words are clustered together and supported by a corpus of relationships. In deep learning frameworks like TensorFlow and Keras, this stage is frequently handled by an embedding layer that keeps a lookup table to translate words

represented by numeric indexes to their dense vector representations.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, None, 128)         9438592

 bidirectional (Bidirectiona  (None, None, 256)        263168
 l)

 bidirectional_1 (Bidirectio  (None, 128)              164352
 nal)

 dense (Dense)               (None, 64)                8256

 dense_1 (Dense)             (None, 64)                4160

 dense_2 (Dense)             (None, 1)                 65

=================================================================
Total params: 9,878,593
Trainable params: 9,878,593
Non-trainable params: 0
_________________________________________________________________
```

Fig. 5.23 layers of the model.

5.3 Implementation of Algorithm with Screen Shots/ Figures

These graphs display the accuracy and loss of the training data sets and accuracy of the testing data sets.

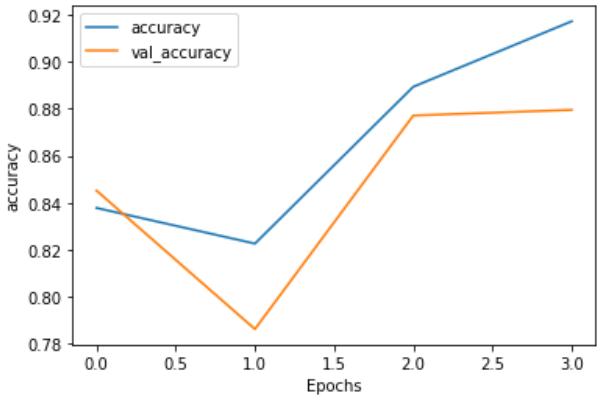


Fig. 5.3a Graph Representing accuracy.

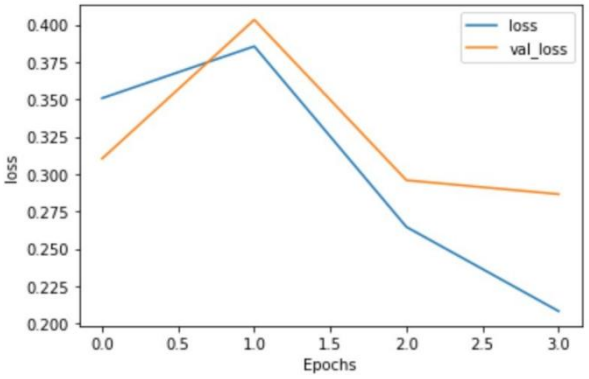


Fig. 5.3b Graph Representing loss.

We've constructed a successful Python sentiment analysis model. We created a binary text classifier in this machine learning project to categorise the reviews' sentiment as either positive or negative. More than 94% accuracy was attained.

```
Enter the text: the product was not good
[22467, 71871, 68611, 6348, 52902]
[22467, 71871, 68611, 6348, 52902, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
1/1 [==============================] - 0s 22ms/step
Sentiment: Negative
[[-2.062323]]
```

Chapter 6

Results

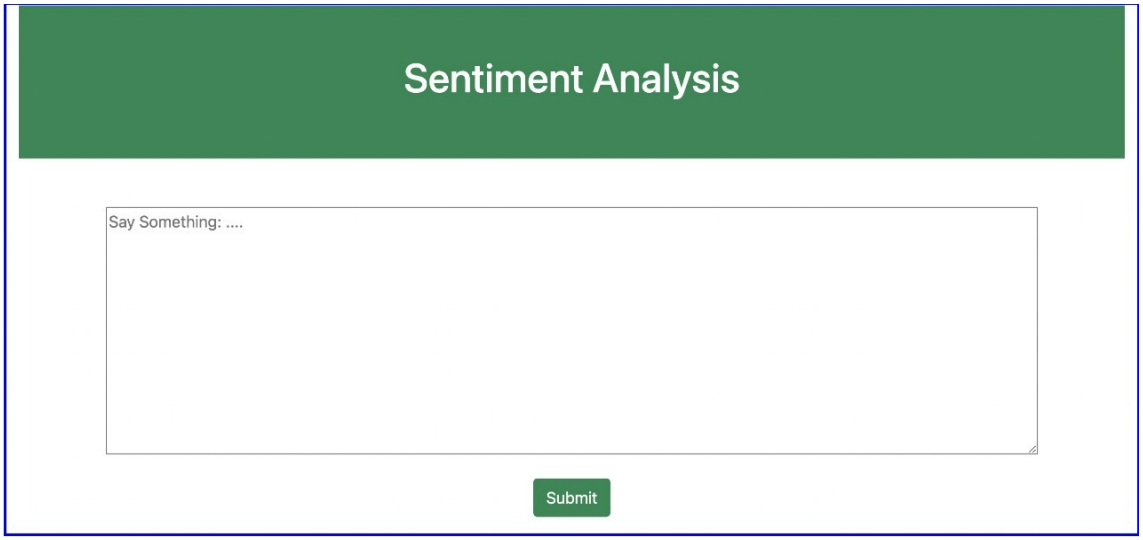
Output:

```
pred_text = input("Enter the text: ")
predictions = predict_fn(pred_text)
if(predictions < 0):
    print("Sentiment: Negative")
if(predictions > 0):
    print("Sentiment: Positive")

print(predictions)
```

```
Enter the text: the product was not good
[22467, 71871, 68611, 6348, 52902]
[22467, 71871, 68611, 6348, 52902, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
1/1 [==============================] - 0s 22ms/step
Sentiment: Negative
[[-2.062323]]
```

This is the website we have done using HTML and CSS.



Chapter 7

Conclusion and Future Scope

7.1 Conclusion

To determine people's opinions, attitudes, and emotional states, sentiment analysis is employed. People may hold either good or unfavorable opinions. common components of Speech is utilized as a feature to derive the text's emotion. Adjective has a role to vital function in separating sentiment from different speech segments. Occasionally, words have When an adjective and an adverb are combined, it can be challenging to determine the sentiment and opinion.

The suggested system extracts the Twitter data in order to do the sentiment analysis of tweets made by the user. The system can calculate each term's frequency as well. To get the findings, a supervised machine learning technique is used.

7.2 Future Scope

- Sentiment analysis will advance to include and comprehend the significance of social media conversations and what they reveal about consumers in the future, going beyond the idea of the quantity of likes, comments, and shares on a post.
- As a result, in order for these businesses to survive in such a cutthroat market, sentiment analysis tools like Bytes View are turning into a necessity.
- In the future we also plan to integrate our model with social media platforms as twitter.

Chapter 8

Bibliography

[1] Chinatsu Aone, Mila Ramos-Santacruz, and William J. Niehaus. AssentorR: An NLP-Based Solution to E-mail Monitoring. In Proceedings of AAAI/IAAI 2000, pages 945--950. 2000.

[2] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79--86. 2002.

[3] Chris Manning and Hinrich Schutze. Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA. 1999.

[4] Satoshi Morinaga, Kenji Yamanishi, Kenji Tateishi, Toshikazu Fukushima. Mining Product Reputations on the Web. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pages 341--349. 2002.

[5] https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17

[6] https://medium.com/analytics-vidhya/natural-language-processing-from-basics-to-using-rnn-and-lstm-ef6779e4ae66

[7] http://cs229.stanford.edu/proj2017/final-reports/5163147.pdf

[8] researchgate.net/publication/341873850_Text_based_Sentiment_Analysis_using_LSTM

[9] https://www.javatpoint.com/epoch-in-machine-learning