MIT | Academy of Engineering

(An Autonomous Institute Affiliated to Savitribai Phule Pune University)

SENTIMENT ANALYSIS OF IMDB MOVIE REVIEW

TY B.Tech. Mini Project Report

SUBMITTED BY

Jayshri Sable	**[T204009]**
Srushti Naikare	**[T204119]**
Manish Shingare	**[T208061]**

GUIDED BY

Prof. Mrs.Neha Hajare

SCHOOL OF COMPUTER ENGINEERING AND TECHNOLOGY,

MIT ACADEMY OF ENGINEERING, ALANDI (D), PUNE-412105

MAHARASHTRA (INDIA)

MAY, 2021

MIT | Academy of Engineering

(An Autonomous Institute Affiliated to Savitribai Phule Pune University)

SENTIMENT ANALYSIS OF IMDB MOVIE REVIEW

TY B.Tech. Mini Project Report

submitted in partial fulfilment of the
requirements for the award of the degree

of

Bachelor of Engineering

in

Computer Engineering & Technology

BY

Jayshri Sable, Srushti Naikare, Manish Shinagare

SCHOOL OF COMPUTER ENGINEERING AND TECHNOLOGY

MIT ACADEMY OF ENGINEERING, ALANDI(D), PUNE-412105

MAHARASHTRA (INDIA)

MAY, 2020

MIT | Academy of Engineering

(An Autonomous Institute Affiliated to Savitribai Phule Pune University)

CERTIFICATE

It is hereby certified that the work which is being presented in the TY B.Tech. Mini Project Report entitled **"*Sentiment Analysis Of IMDb Movie Review*",** in partial fulfillment of the requirements for the award of the **Bachelor of Technology in Computer Engineering & Technology** and submitted to the **School of Computer Engineering & Technology, Alandi(D), Pune, Affiliated to Savitribai Phule Pune University (SPPU), Pune** is an authentic record of work carried out during an Academic Year 2020-2021, under the supervision of **Mrs.Neha Hajare , School of Computer Engineering & Technology.**

Jayshri Sable	**PRN No. 0120180073**	**Exam Seat No. T204009**
Srushti Naikare	**PRN No. 0120180565**	**Exam Seat No. T204119**
Manish Shingare	**PRN No. 0120180568**	**Exam Seat No. T208061**

Date:

Signature of Project Advisor

Mrs. Neha Hajare

School of Computer Engineering &Technology,

MIT Academy of Engineering, Alandi(D), Pune

Signature of Dean

Mrs. Ranjana Badre

School of Computer Engineering &Technology,

MIT Academy of Engineering, Alandi(D), Pune

(STAMP/SEAL)

Signature of Internal examiner/s

Name……

Affiliation……

Signature of External examiner/s

Name……

Affiliation……

ACKNOWLEDGEMENT

We want to express our gratitude towards our respected project guide Prof Mrs. Neha Hajare under whom we have carried our project work. Her constant encouragement and valuable guidance during the project work encourage us with constant flow of energy to continue the work.

We also want to express our gratitude towards respected School Dean Mrs. Ranjana Badre for her continuous encouragement.

We would be failing in our duty if we do not thank all the other staff and faculty members for their experienced advice and evergreen co-operation

1. Jayshri Sable
2. Srushti Naikare
3. Manish Shingare

ABSTRACT

Sentimental Analysis is a new variant in the research area. It basically refers to opinions or views of the different data that is being collected using surveys , comments and reviews over the web. The data set we have used in our project is from imdb movie rating website. Through this we can classify the review as positive or negative, a large amount of data is being generated daily. This will determine the polarity of reviews and these reviews will be classified as three types positive, negative and neutral.

LIST OF FIGURES

CONTENTS

1.INTRODUCTION

What is Sentimental Analysis?

Sentiment analysis refers to the use of natural language processing, text analysis and computational linguistics to extract and identify subjective information in the whole sentence. It aims to determine the attitude of the speaker or writer with respect to some topic, in our case, with respect to movies. The attitude can be his or her judgement or evaluation. Affective state (i.e. emotional effect of writer when writing).The intended emotional communication (i.e. emotional effect the writer wishes to have on the reader).
In the last few years there has been a rise in usage of social media such as blogs and social media networks such as Instagram, twitter, etc. which has fueled the interest in sentiment analysis.

Online opinion has turned into a kind of virtual currency with rapid increase of reviews, ratings, recommendations and other forms of expressions, for any business that is looking to market their products, identify new opportunities and manage their reputation. In order to automate the process of filtering the noise, understanding the conversations, identifying customers or viewers needs, many are now looking to the field of sentiment analysis. But the main problem of most of the sentiment analysis algorithms is that they use simple terms to express sentiment about a product or service. However, cultural factors, sentence negation, sarcasm, language ambiguity, use of slangs and differing context make it very difficult to turn the string of text into simple pro or con sentiment.

A fundamental task in sentiment analysis is classifying the polarity of given text. It focuses on whether the expressed opinion is positive, negative or neutral. Sometimes it goes beyond polarity and looks at emotional states such as “angry”, “sad” and “happy”.
Existing approaches to sentiment analysis can be grouped into four main categories. They are keyword spotting, statistical methods, and concept-level techniques. Keyword spotting classifies text by affect categories based on the presence of unambiguous affect words such as happy, sad, afraid, and bored. Statistical methods influence elements from machine learning such as latent semantic analysis, support vector machines, bag of words and Semantic Orientation. Concept-level approaches through the analysis of concepts that do not explicitly convey relevant information, but which are implicitly linked to other concepts that do so.

So our project, unlike others, aims to do sentiment analysis using lexical approach. In most of the string of text or sentences expressing the opinion about the movie, their thoughts which contain different writing habits, use of slang words, sarcasm, non-english words etc. so using lexical approach and using different libraries, we can easily counter those problems. Making use of tokenization our system would be easily identifying the main keywords in the sentences.

Then the sentence is examined for subjectivity, only the sentence with subjective expressions are retained and sentences which convey objective expression are discarded. In this manner the filtration and text preparation would do the rest work easy. Then using some common computation techniques sentiment detection would be done. Sentiment classification i.e. identification of the given sentiment is either positive or negative, is done

1.1 Motivations

Motivation for the project:- In the era of technology each and everything is going online. From as simple as booking bus or flight tickets to as complex as trading online. Every business is having their own websites on internet to provide different services such as, buy monthly grocery, make payments, book tickets, watch online movies etc. For these businesses, customers are god like. Customer's feedback / review/ opinions are very important to businesses, to analyse their growth, take closer look at their market, make changes accordingly and stand on the customer's demands.

Even industry like film industry, film makers are so much dependent on customers/ consumers that they constantly need to keep a close eye on viewers likes/ dislikes , opinions, feedbacks to make appropriate decisions while making films. This is where our project, Sentiment Analysis of Movie reviews kicks in.

1.21.2 Problem Statement

To perform Sentiment Analysis on IMdb reviews of Movies by lexicon approach and to identify whether review is positive, negative or neutral.

1.31.3 Objectives and Scope

To implement an algorithm for automatically classification of text into positive, negative and neutral.

To study and implement tokenization.

To study and implement lemmatization.

To calculate polarity value of review.

To calculate accuracy of the algorithm.

2. LITERATURE SURVEY

[1] Walaa Medhat of a School of Electronic Engineering, Canadian International College, Cairo Campus of CBU, Egypt has published (Ain Shams University) a paper based on Sentiment analysis algorithms and applications which can three main classification levels in SA: document-level, sentence-level, and aspect-level SA. They are presented an overview on the recent updates in SA algorithms and applications. s. After analyzing these articles, it is clear that the enhancements of SC and FS algorithms are still an open field for research.

The entity can represent individuals, events or topics. These topics are most likely to be covered by reviews. The two expressions SA or OM are interchangeable. These articles give contributions to many SA related fields that use SA techniques for various real-world applications. After analyzing these articles, it is clear that the enhancements of SC and FS algorithms are still an open field for research. They express a mutual meaning. They only represent the level on it so we are going for the next paper. Sentence-level SA aims to classify sentiment expressed in each sentence.

Sentiment Classification techniques can be roughly divided into machine learning approach, lexicon based approach and hybrid approach [69]. The Machine Learning Approach (ML) applies the famous ML algorithms and uses linguistic features. The Lexicon-based Approach relies on a sentiment lexicon, a collection of known and precompiled sentiment terms.

[2] Maite Taboada of a Simon Fraser University has published (Simon Fraser University) A paper based on Semantic orientation (SO) is a measure of subjectivity and opinion in text. It usually captures an evaluative factor (positive or negative).From these they are presented a word-based method for extracting sentiment from texts. This attribute to this criteria for selecting and ranking words, which include excluding ambiguous words and including fewer rather than more words. There exist two main approaches to the problem of extracting sentiment automatically. The lexicon-based approach involves calculating orientation for a document from the semantic orientation of words or phrases in the document. We have presented a word-based method for extracting sentiment from texts. Building on previous research that made use of adjectives, we extend the Semantic Orientation Calculator to other parts of speech.

We also introduce intensifiers, and refine our approach to negation. SO-CAL is applied to the polarity classification task, the process of assigning a positive or negative label to a text that captures the text's opinion towards its main subject matter. We show that SO-CAL's performance is consistent across domains and on completely unseen data.

The majority of the statistical text classification research builds Support Vector Machine classifiers, trained on a particular data set using features such as unigrams or bigrams, and with

or without part-of-speech labels, although the most successful features seem to be basic unigrams

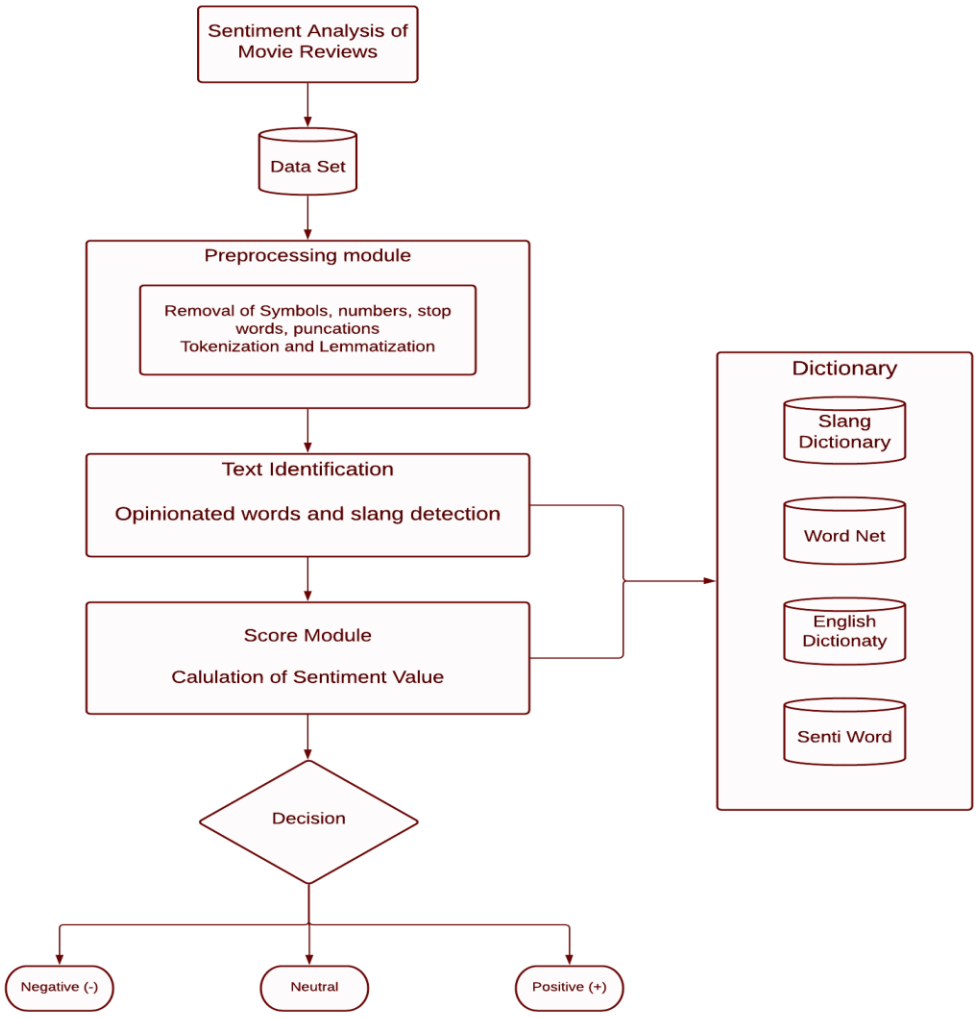
[3] Karthik Konar 1MCA Student, Dept. of Computer Engineering, NMIMS Mukesh Patel School of Technology Management & Engineering, Vile Parle(West) Mumbai has published (NMIMS Mukesh Patel School of Technology Management & Engineering, Vile Parle(West) Mumbai) in their paper which can be based on Sentimental analysis is that particular domain, where you try to understand human emotions with the help of a software. Human emotions are in written form, and we classify those sentiments as positive, negative, neutral. This paper provides a detailed comparison of various applications of sentiment analysis which was implemented using different approaches such as lexicon, machine learning, and VADER sentiment analysis.

There are various applications of sentimental analysis such as Review classification, Product review mining, etc. In review classification, we classify the user reviews into 3 categories i.e. positive, negative, and neutral. For e.g.: There are a lot of customers, who post a lot of reviews, how do we know the sentiment of each customer? This is where we can use sentimental analysis and classify the reviews to be positive, negative, neutral. Based on the experiment performed we found that VADER sentiment analysis provides the highest accuracy when compared with the lexicon and machine learning approach.

VADER Sentiment analysis approach to classify the accuracy of positive and negative reviews and the time taken for execution is very less when compared with the lexicon and machine learning approach.

3. SYSTEM DESIGN

3.1 Block Diagram



(Figure 3.1 Block Diagram)

For our project we have proposed the block diagram that explains how the data undergoes the execution in the whole program along with detailed explanation .

Dataset

It contains a record of reviews of different movies. It which contains over 5000 reviews with their respective sentiment for future analysis

Preprocessing Module

In this section preprocessing steps are carried out on a dataset. It includes data cleaning, tokenization, lemmatization which is explained later on in detail.

Text Identification

This module performs tasks where words are identified and classified with the help of dictionaries.

Dictionary

It contains sentiment resources which helps for Text Identification. These sentiment resources(dictionaries) contain positive and negative sentiment words. Some misspelled words are also included in the lexicon as they appear frequently in the social media text.

Score Module

It will help us to calculate the sentiment value of the review with the help of a dictionary which has sentiment values for words. It calculates positive, negative and neutral sentiment value a review separately and then it would calculate the final sentiment score

EXPLANATION :

1. Here the dataset is first imported from the database

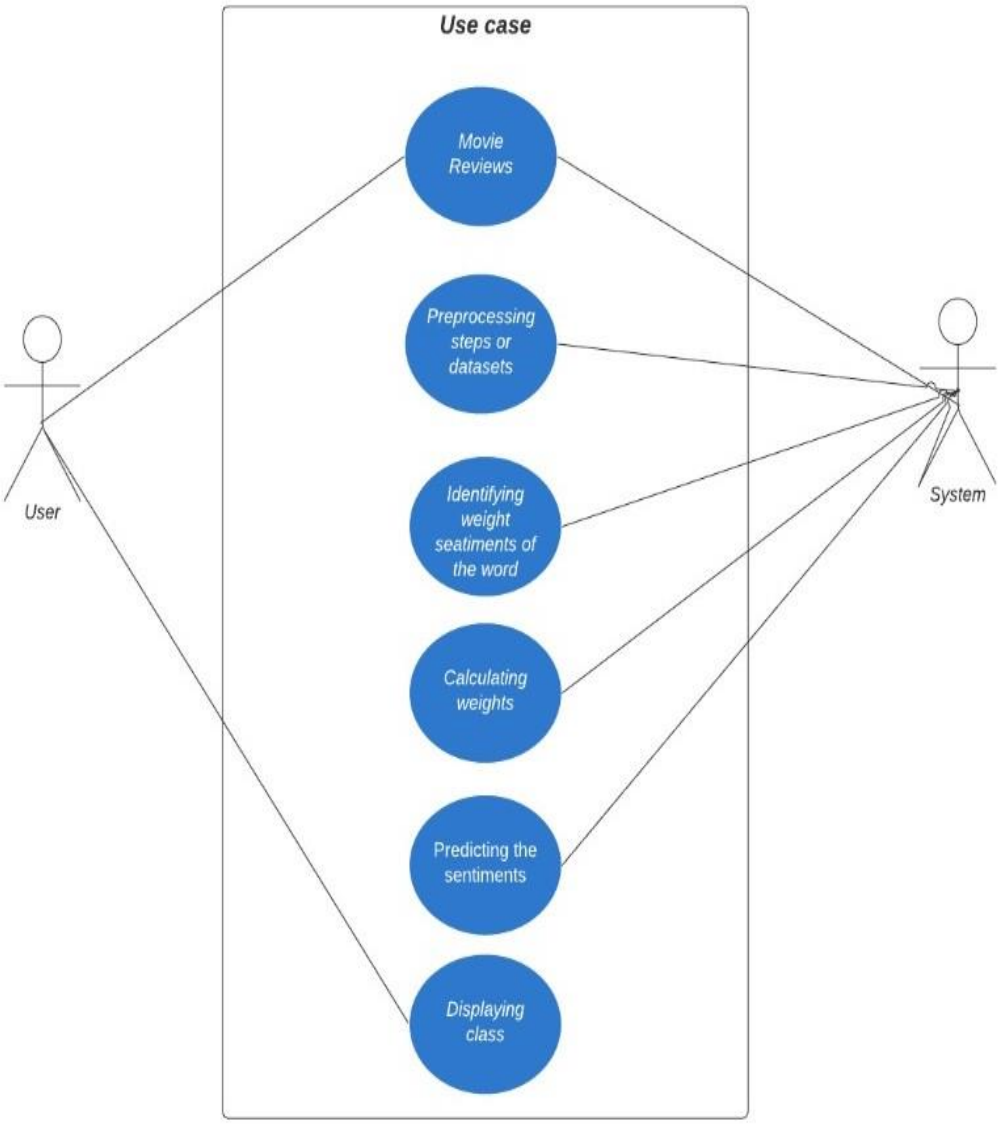
2. Preprocessing steps are carried out on a dataset. It includes data cleaning, tokenization, lemmatization

3. It identifies the words based on adjectives and classified with the help of dictionaries.

4. Then with the help of score module, calculate sentiment value

5. The review will be classified into positive, negative, neutral based on sentiment value

3.2 Use case diagram

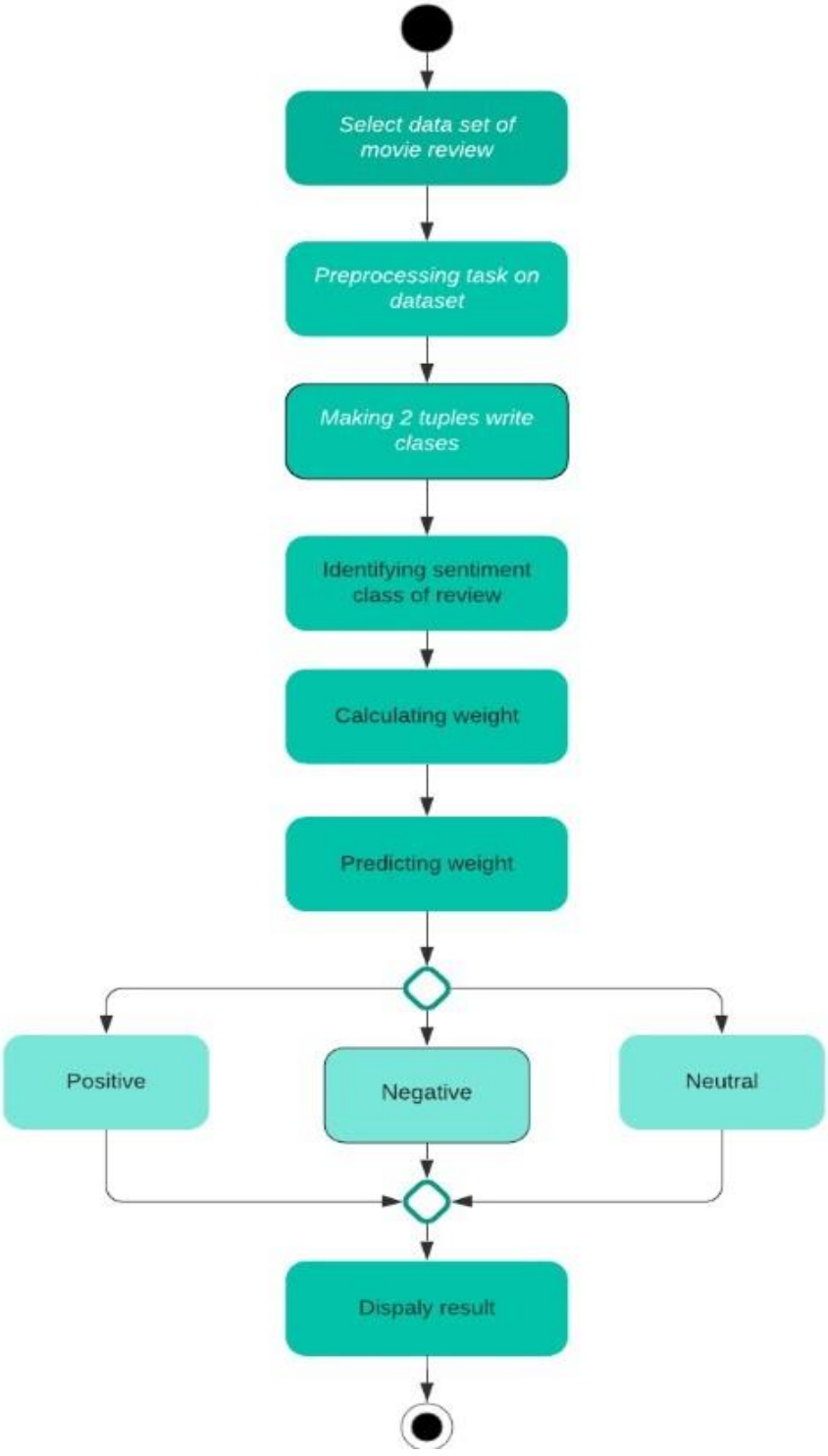


(Figure 3.2 Use case Diagram)

3.3 Sequence diagram

(Figure 3.3 Sequence Diagram)

3.4 Activity diagram



(Figure 3.4 Activity Diagram)

3.5 Hardware & Software requirement:

- **Windows 10 Machine**

(8 GB RAM, 256 SSD)

- **Anaconda (Virtual Environment)**
- **Jupyter Notebook**

4. IMPLEMENTATION DETAILS

4.1 Algorithm analysis :

Step 1: Importing the data set

Step 2: Removing single letters and numeric from the reviews

Step 3: Snippet to tag the parts of speech for each review

Step 4: Narrow down each review with only adjectives and verb forms

Step 5: Making a list of tuples which has review and its respective class

Step 6: Separating the data set by each class i.e. "Positive", "Negative", "Neutral"

Step 7: Merge the reviews in each class in to three single lists

Step 8: Splitting the words in each list to get the word frequencies

Step 9: Finding the frequencies of the unique words in each class

Step 10: Snippet to set the dictionary size

Step 11: Identifying the similar words for each class in each review

Step 12: Calculating the score of positive, negative and neutral words in each review

Step 13: Predicting the class based on the score

4.2 Result

Creating a data frame with respective weigths

```
In [39]: UL = pd.DataFrame()

In [40]: UL['POS']= PW_1

In [41]: UL['NEG']= NGW_1

In [42]: UL['NEU']= NUW_1

In [43]: UL['truth']=A['sentiment']

In [44]: UL['PRE'] = A['review']

In [45]: UL.head(5)
```

Out[45]:

	POS	NEG	NEU	truth	PRE
0	1.00	0.92	0.0	positive	One of the other reviewers has mentioned that ...
1	1.00	0.81	0.0	positive	A wonderful little production. The...
2	1.00	0.86	0.0	positive	I thought this was a wonderful way to spend ti...
3	0.72	1.00	0.0	negative	Basically there's a family where a little boy ...
4	1.00	0.76	0.0	positive	Petter Mattei's "Love in the Time of Money" is...

(Figure 4.1 Result 1)

Predicting the class based the weights.

```
In [46]: for i in range(500):
             if (UL.POS[i]-UL.NEG[i] > 0.03):
                 UL.PRE[i] = 'positive'
             if (UL.POS[i]-UL.NEG[i] < -0.03):
                 UL.PRE[i] = 'negative'
             elif (UL.POS[i]-UL.NEG[i] <= 0.03 and UL.POS[i]-UL.NEG[i] >= -0.03):
                 UL.PRE[i] = 'neutral'
```

```
F:\anaconda\lib\site-packages\ipykernel_launcher.py:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  This is separate from the ipykernel package so we can avoid doing imports until
F:\anaconda\lib\site-packages\ipykernel_launcher.py:5: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """
F:\anaconda\lib\site-packages\ipykernel_launcher.py:7: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  import sys
```

```
In [47]: UL.head(5)
```

Out[47]:

	POS	NEG	NEU	truth	PRE
0	1.00	0.92	0.0	positive	positive
1	1.00	0.81	0.0	positive	positive
2	1.00	0.86	0.0	positive	positive
3	0.72	1.00	0.0	negative	negative
4	1.00	0.76	0.0	positive	positive

(Figure 4.2 Result 2)

```
In [48]: pd.crosstab(UL.truth,UL.PRE)
```

Out[48]:

truth \ PRE	negative	neutral	positive
negative	257	6	0
positive	0	14	223

```
In [49]: print (accuracy_score (UL.truth,UL.PRE))
```

```
0.96
```

```
In [50]: print (recall_score(UL.truth,UL.PRE, pos_label= 'negative', average='micro'))
```

```
0.96

F:\anaconda\lib\site-packages\sklearn\metrics\_classification.py:1321: UserWarning: Note that pos_label (set to 'negative') is
ignored when average != 'binary' (got 'micro'). You may use labels=[pos_label] to specify a single positive class.
  % (pos_label, average), UserWarning)
```

```
In [51]: UL.to_csv("ultimate.csv", sep=',', encoding='utf-8')
```

(Figure 4.3 Result 3)

```
In [1]: import pandas as pd
```

```
In [10]: import numpy as np
         import pandas as pd
         import matplotlib.pyplot as plt
```

```
In [4]: A = pd.read_csv('C:/Users/Mangesh Shingare/ultimate.csv')
```

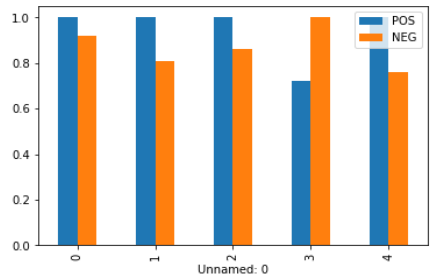
```
In [5]: A.head()
```

Out[5]:

	Unnamed: 0	POS	NEG	NEU	truth	PRE
0	0	1.00	0.92	0.0	positive	positive
1	1	1.00	0.81	0.0	positive	positive
2	2	1.00	0.86	0.0	positive	positive
3	3	0.72	1.00	0.0	negative	negative
4	4	1.00	0.76	0.0	positive	positive

```
In [22]: A.head().plot.bar(y=['POS','NEG'], x='Unnamed: 0')
```

```
Out[22]: <matplotlib.axes._subplots.AxesSubplot at 0x190b5198a88>
```



(Figure 4.4 Result 4)

5. CONCLUSION & FUTURE SCOPE

This paper about Sentiment Analysis of IMDb Movie Reviews using Lexicon approach. We have presented a Dictionary-based method for extracting sentiment from texts The algorithm provides a clear idea about the importance and working of the Sentiment predictions. It provides accuracy up to 96%

Sentiment analysis is a uniquely powerful tool for businesses that are looking to measure attitudes, feelings and emotions regarding their brand. To date, the majority of sentiment analysis projects have been conducted almost exclusively by companies and brands through the use of social media data, survey responses and other hubs of user-generated content.

The future of sentiment analysis is going to continue to dig deeper, far past the surface of the number of likes, comments and shares, and aim to reach, and truly understand, the significance of social media interactions and what they tell us about the consumers behind the screens.

Our project tries to identify the sentiments value of the review classifying it into positive or negative review. Later on we can create GUI interface which make ease the task of classifying the review.

REFERENCES

[1] Walaa Medhat , Ahmed Hassan , Hoda Korashy “Sentiment analysis algorithms and applications: A survey ” School of Electronic Engineering, Canadian International College, Cairo Campus of CBU, Egypt b Ain Shams University, Faculty of Engineering, Computers & Systems Department, Egypt Issued: 19 | April 2014

[2] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Vol, Manfred Stede “ Lexicon-Based Methods for Sentiment Analysis ” Department of Computer Science, University of British Columbia, 201-2366 Main Mall, Vancouver, B.C. V6T 1Z4 Canada. Issue: 28 | September 2010.

[3] Karthik Konar , Riddhi Patel “ A Systematic Study on Sentimental Analysis and its Applications ” MCA Student, Dept. of Computer Engineering, NMIMS Mukesh Patel School of Technology Management & Engineering, Vile Parle(West) Mumbai Issue: 06 | June 2020

[4] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau Department of Computer Science Columbia University New York, NY 10027 USA Issued: 23 | June 2011

[5] Fazal Masud Kundi, Aurangzeb Khan, Shakeel Ahmad, Muhammad Zubair Asghar “Lexicon-Based4 Sentiment Analysis in the Social Web ” Institute of Computing and Information Technology, Gomal University, D.I.Khan, Pakistan, Institute of Engineering and Computer Sciences, University of Science and Technology Bannu, Pakistan Issued: 26 | May 2014

[6] Sentiment Analysis: Types, Tools, and Use Cases (https://www.altexsoft.com/blog/business/sentiment-analysis-types-tools-and-use-cases/)

[7] Dhiraj Murthy, Twitter and elections: are tweets, predictive, reactive, or a form of buzz?, Information, Communication & Society, 18:7, 816-831, DOI:10.1080/1369118X.2015.1006659

[8] Humera Shaziya, G.Kavitha, Raniah Zaheer, 2015, Text Categorization of Movie Reviews for Sentiment Analysis , International Journal of Innovative Research in Science, Engineering and Technology, Vol. 4, Issue 11.

[9] Akshay Amolik, Niketan Jivane, Mahavir Bhandari, Dr .M. Venkatesan, Twitter Sentiment Analysis of Movie Reviews using Machine Learning Techniques, School of Computer Science and Engineering, VIT University, Vellore.

[10] Palak Baid,Apoorva Gupta ,Neelam Chaplot , Sentiment Analysis of Movie Reviews using Machine Learning Techniques, International Journal of Computer Applications (0975 – 8887) Volume 179 – No.7, December 2017 45