Project Report

on

“Opinion mining of public reviews”

Submitted as Mini Project Report

FOR MINI PROJECT LAB(KCS-554)

Session 2022-23

in

Computer science and engineering

By
Karan Singh
Nitin Singh

Under the guidance of

Mr. Abhishek Kumar Shukla

ABES ENGINEERING COLLEGE, GHAZIABAD

AFFILIATED TO
DR. A.P.J. ABDUL KALAM TECHNICAL UNIVERSITY, U.P., LUCKNOW
(Formerly UPTU)

STUDENT'S DECLARATION

I / We hereby declare that the work being presented in this report entitled **"Opinion mining of public review"** is an authentic record of my / our work carried out under the supervision of **"Mr. Abhishek Shukla".**

The matter embodied in this report has not been submitted by me/us for the award of any other degree.

Dated: **Signature of Students**

Karan Singh

Roll no:2000320100078

Nitin Singh

Roll no:2000320100109

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

.

Signature of HOD

Name: Prof. Divya Mishra

CSE

Date...........................

ACKNOWLEDGEMENT

It gives us great pleasure to present the report of the B. Tech Mini Project undertaken during B. Tech. Third Year. We owe special gratitude to **Mr. Abhishek Kumar Shukla** *for his constant support and guidance throughout our work. Her sincerity, thoroughness, and perseverance have been a constant source of inspiration for us. It is only his cognizant efforts that our endeavors have seen the light of day.*

We also take the opportunity to acknowledge the contribution of Professor **Dr. Divya Mishra***, Head, Department of* **CSE***, ABESEC Ghaziabad for his full support and assistance during the development of the project.*

We also do not like to miss the opportunity to acknowledge the contribution of all department faculty members for their kind assistance and cooperation during the development of our project. Last but not the least, we acknowledge our friends for their contribution to the completion of the project.

Signature:

Name:

Roll No.:

Date :

ABSTRACT

Sentiment analysis plays a very important role in our lives. Mining of opinions are very crucial in all fields from e commerce websites to social media sites. Mining of opinions can be extensively used in the fields where opinions play

a major role. The products on any e commerce websites have thousands of reviews which helps customers to buy a product. Social media websites also have people with large number of opinions on a particular subject. This project caters to this need and classifies the opinions of people as positive and negative. This can further be used by movie recommendation systems and e commerce websites for evaluation of their product. The problem statement involved in this project is to classify the opinions as positive and negative with the help of deep learning algorithms by achieving high accuracy. The procedures involved in this project will be of dataset selection, data preprocessing, data tokenization, data cleansing and building a neural network. We have taken the dataset of product reviews for this purpose. Data preprocessing and data cleansing is done so that deep learning algorithms can be easily applied on the data. Sentence Tokenization is used so that it splits the content into particular sentences. Word Tokenization is done to split the content into particular words. Deep learning algorithms learn on their own and do not require guidance. The main objective of using deep learning model is for increasing efficiency, performance and accuracy. This project classifies the opinions with maximum efficiency.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

SERIAL NO. **FIGURE NO. CAPTION** **PAGE NO.**

LIST OF ABBREVIATIONS

ROC	Receiver Operating Characteristics
SA	Sentiment Analysis
CNN	Convolutional neural network
RNN	Recurrent neural network
SNN	Simple neural network

CHAPTER 1 INTRODUCTION

1.1. Problem Introduction:

The world of internet has brought us everything at ease, from buying products to searching the next new product purchases. When you have a platform to host a great website to promote the products which you have, the internet is now turning into a place where consumers check out products and explore things. Then they take decisions accordingly.

It is very important to imagine the value of customer opinions, but there is nothing to highlight beyond the data in terms of how reviews are used and how they affect the business. Statistical methods analyze how customers behave prior to and after using services or purchasing products, which can help them improve business development plans.

Reviews can not only have the capability to affect consumer decisions but also build up the company's image. Reviews have the potential to win customer feedback, and encourage people to engage with the company. Customer communication ultimately leads to improved business benefits. Reviews have utmost importance in our daily lives from e-commerce websites, social media platforms to political reviews.

Classification of reviews as positive and negative is the need of the hour. The machine learning algorithms and tools can provide information by analyzing product reviews automatically and categorizing it by positive, neutral, and negative. We need a review sentiment analyzer system which can accurately do this task for us.

This project will cater the need to classify the reviews in an efficient manner.

By using sentiment analysis to organize product reviews, we can understand our customer’s likes and dislikes for our product, can compare our product reviews with competitors, and can get the latest product details.

1.1.1. Motivation

According to consumer's vision:

While making a decision it is very important that we know the opinion of the people around us. Earlier the group was usually small, with a few loyal friends and family members. But, now with the arrival of the Internet we see people expressing their opinions on blogs and forums.

These are now being studied diligently by people who want an idea about a particular business (product, movie, etc.).

Therefore, there are a lot of ideas available on the Internet. From a consumer perspective, getting ideas about a particular business is important. To try transferring such a large amount of data to understand the general view is impossible users with a large amount of this data.

Therefore, there is a need for a system for positive reviews and negative reviews. In addition, writing these articles about their feelings will provide a brief summary of students with a general view of the business.

According to manufacturer's vision:

With the explosion of Web 2.0 platforms like blogs, forums, etc., Consumers have a platform for sharing their product knowledge and their ideas, good or bad in any way product or service.

According to Pang and Lee (2008) these words of consumers can have a profound effect in building the opinions of other consumers and, ultimately, their product integrity, their purchasing decisions, and their product representative.

As consumers begin to use the power of the Internet to increase their horizons, it has become an explosion of review sites and blogs, where users can see the benefits of a product or service and errors.

These ideas therefore shape the future of a product or service. Vendors need a system which can identify styles in customer reviews and use them to improve their product or service as well as identify future needs.

According to community's Vision:

Recently, certain events, affecting the Government, have been caused by the use of the Internet. Social networks are used to bring people together to organize mass gatherings and to oppose oppression.

On the black side, social networks are used to suggest people about race or the human race, which has led to massive loss of life. Therefore, there is a need for Sentiment Analysis systems can identify such items and reduce them if necessary.

1.1.2. Project Objective:

Sentiment analysis basically involves analyzing the information under a text and classifying the polarity of the opinion according to negative and positive paraameter. In decision making, the preconditioning of others have a major impact on the convenience of customers for taking main decisions about online purchases. Sentiment is one of the many areas of computer studies that deal with natural language-based analysis. Such theoretical studies include, among other things, classification, emotional recognition and emotional impact, quality, value calculation, ideas in the text, identifying the source of the text and summarizing the concepts. Sentiment analysis has emerged as an exciting new trend in social media with a wide range of active applications ranging from business programs (intelligent marketing, benchmarking and benchmark performance and optimization), applications such as subcontracting technology. Sentiment Analysis, the site of Natural Language Processing (NLP), is used to classify reviews using the sense of words to be classified as positive or negative. Using the sense expressed in words or text, ideas in any entity can be divided into positive or negative. For example, the phrase, ‘I am not happy with this product even though it is very cheap’ expresses negative feelings about this product. The level of feeling used is also considered. For example, 'I like this product' shows a much better feeling than the sentence 'I like this product'. Apart from the common adjectives such as 'good', 'bad' and 'very good', conjunctions such as 'but', 'although', 'while' also have a voice in the full view of the sentence. Many challenges as organizations and individuals try to analyze and understand the opinions of others. Unfortunately to find sources of information monitor and analyze it are herculean activities. It is impossible to manually find sources online, extract ideas from them and express them in a common format. In recent years, millions of people express unresolved opinions about the various product features and their nuances. This creates an effective response that is important not only for product development companies, but also for competitors and other potential customers. Hence, the main objective of our project is to meet the above requirements and classify the reviews, sentiments or opinion as positive or negative. It uses neural networks and deep learning algorithms that learn on their own and do not require guidance and classifies the reviews accurately

1.1.3. Scope of the Project

In today’s scenario, social platforms are shooting up; the vast data can be used to meet business objectives, marketing, and other promotional strategies for their profits. The benefit of social media to mine public opinions and analyze their emotions can be obtained by opinion mining techniques.

Major scope of opinion mining include: -

- By using sentiment analysis different customer segments of your business analysis gets easy and helps us to have a better understanding of sentiment and opinions of people.
- Opinion mining can be used for analyzing political opinions.
- In stock market analysis.
- Movie recommendation systems.

1.2. Related Previous Work

Sentiment Analysis and classification in the field of Machine learning can be performed in two ways. One method is supervised learning and another is unsupervised learning. Support vector machines (SVM) and naïve Bayes (NB) widely use monitoring techniques. Solutions associated with machine learning solutions involve the construction of separators from a documents, where we can represent each text as a bag of words . Also, it is common to use certain methods to prevent and eliminate word loss. In general, the ethical categories in the domain in which they are trained do not reflect the same behavior in another domain because they rely heavily on the training data used.

Many research papers use machine learning techniques along with LDA Analysis on naive bayes. Unsupervised machine learning technique is also used for this purpose. The machine learning algorithm is used to categorize the positive and negative reviews. Preprocessing is also done on the unstructured data. Building the vocabulary and extracting the features are a very important part of the process. This research work uses naïve bayes algorithm because of its high accuracy. Sentiment analysis can be used in a lot of decision making areas also. Sentiment Vader (VALENCE AWARE DICTIONARY AND SENTIMENT REASONER) is sentiment analysis instrument that is used to analyze social media data. Senti word net is used for natural language processing tool to calculate frequency and significance of word.

As we know that sentiment analysis is one of the most tedious and difficult task in the application of natural language because people also strive to analyze emotions accurately and the process of extraction of the features and classification is not so efficient and the model does not get the desired accuracy. To overcome this, we have tried to implement a sentiment analysis model using Deep Neural Networks that performs better and give better accuracy and classifies almost every review correctly. The initial training of a deep learning model was extremely time-consuming and often required millions of data points until it began to learn on its own. But with deep learning the model gets trained and it performs far better than the machine learning model and gets a better efficiency and accuracy. Feature extraction gets very efficient and faster by using the deep neural networks.

1.3. Organization of the Report.

I. **Introduction:** Reviews have utmost importance in our daily life whether it be ecommerce websites, Social media platforms or political reviews. Classification of reviews as positive and negative is the need of the hour. We need opinion mining of public reviews for this classification. This model will cater the need to classify the reviews in efficient manner. Sentiment analysis refers to the natural language processing which deals with the classification of opinions and emotions expressed in written. Opinion Mining aims at studying public opinions, views, and emotions towards any area be it an individual, any content, any incident or product.

2. **Literature Survey:** This section explains some of the research works in the field of Machine Learning and data mining in order to analyze opinions and preparing prediction model for various applications. Multiple algorithms and methods are used

   to find the influence of reviews or opinions onthe working of a system. Views are mined only at keywords level rather than whole texts poste by the customers. Majorly unsupervised learning approach is used to catch patterns and cohesion in reviews.

3. **Methodologies:**

•Data pre-processing :Removal of duplicate and repeated words, stop-words, URLs, usernames and conjunctions which do not contribute in sentiment is pre-processing or stemming.

- Machine learning algorithms: Different machine learning algorithms like Naïve Bayes, Support Vector Machine(SVM) etc are used. Then neural network algorithms like CNN,RNN and SNN are used and their accuracy is compared.

- Feature Extraction: Some words can take different meaning in different sentences so POS tagging is used.

CHAPTER 2

LITERATURE SURVEY

2.1 Introduction:

According to [5] Prof. Brian Keith NorAm Buena and his research paper, Sentiment analysis is shooting up these days.

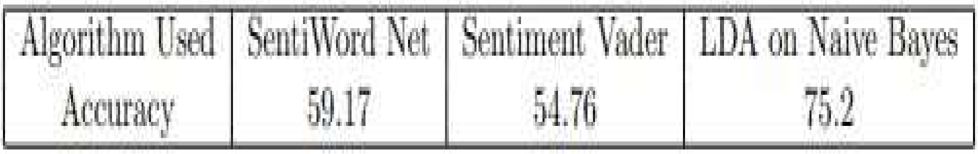
One of the applications of opinion mining is product reviews. This application has its own importance in the market as people views regarding certain product and service matters to achieve maximum profits.

One major area where sentiment analysis is yet to be applied is opinion mining for knowing the aim or alignment of scientific research paper review. Although there are certain hurdles which are predicted to be faced while achieving this application such as unbalanced classes, biasness towards negative reviews, language barriers and listed anonymous reviews.

According to MirsaKarim [1] a customer analysis is done in this research paper. These reviews are conducted to take into account any customer needs where the review plays an important role. Neurological Analysis is an ongoing field of exploration that continues to be important due to the use of various applications. Audit is provided by people in an informal manner in the form of forums, blogs etc. After that the re-analysis is done and to see if the research is valid, it is not good or non-aligned. Order methods such as vocabulary and machine-based methods are used for Opinion Analysis. The wordbased method is in the direction of the word index. Machine learning techniques are very much used for sentiment analysis these days. The system takes a label less database where the test is not visible by its marker.

This research paper has used rule based and Machine learning algorithms and found that machine learning process is very accurate and precise in anticipation of the conclusion of the sentence or feelings of discovery associated with the sentence. Future work can be improved as a study to join the separation in various ways.

Table 2.1.1 Accuracy of rule based mechanism

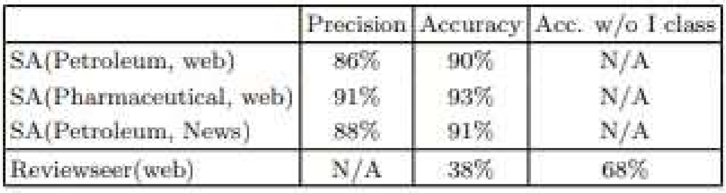


Algorithm Used	SentiWord Net	Sentiment Vader	LDA on Naive Bayes
Accuracy	59.17	54.76	75.2

Sentiment analysis has become a very important part of our lives. According to [3]S.Muthukumaran and A.Victoria Anand Mary, In this research paper sentiment analysis is described as a Natural Language Processing problem, which differentiate opinionated text and categorize it as a negative opinion, a neutral opinion, or a positive one.

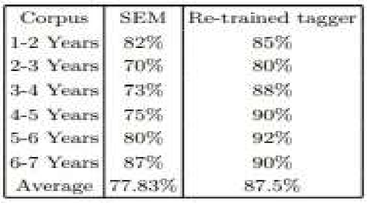
Opinion Mining and Sentiment Analysis, extensions for opinion mining, natural language processing techniques and text analysis which is a process that determines the feelings of people or emotions or attitudes on a particular topic through a great unstructured processing online content. Opinion Mining releases feelings or ideas with words that are present in free text during Sentiment analysis determines polarity, whether positive or negative or neutral by analyzing each word or sentence. Sentiment Analysis throws light on the author's or speaker's point of view on a particular topic and can be done at word or feature level, sentence level and text level.

Table 2.1.2 Characteristics of the initial SA



	Precision	Accuracy	Acc. w/o I class
SA(Petroleum, web)	86%	90%	N/A
SA(Pharmaceutical, web)	91%	93%	N/A
SA(Petroleum, News)	88%	91%	N/A
Reviewseer(web)	N/A	38%	68%

Table 2.1.3 Impact of the re-learning on the accuracy of the distinct age groups



Corpus	SEM	Re-trained tagger
1-2 Years	82%	85%
2-3 Years	70%	80%
3-4 Years	73%	88%
4-5 Years	75%	90%
5-6 Years	80%	92%
6-7 Years	87%	90%
Average	77.83%	87.5%

According to [4] Mohd Ridzwan Yaakub1 and Muhammad Iqbal Abu Latiffi2, sentiment analysis is actually a way of dividing emotions into positive, negative or neutral ones in respect of any product, service or problem. These days, sharing of ideas, shared marketing, online ticket bookings, and online shopping are on the rise in the daily life of the people. Some people's opinions are the most important part of knowledge for most of us. Facebook, Instagram are the social networking sites that allow users to post comments or ideas about any issues and topics.

2.2 Techniques and methodologies:

The techniques used in the process are natural language processing with POS tagging combined with unsupervised methods like Scoring algorithms. This gives a whole sentence of a review a syntactic structure which makes the analysis less complex. Citation from the paper says According to [5]Brian Keith Norambuena and, Exequiel Fuentes Lettura ,A set of tests was performed to test the strength and effectiveness of the proposed methods related to the original base, using standard methods, such as The data set is collection of reviews in an international conference in Spanish evaluated based on 5-point scale that is –2 for ‘extremely negative’, -1 for ‘negative’, 0 for ’neutral’, 1 for positive’, 2 for very- positive.

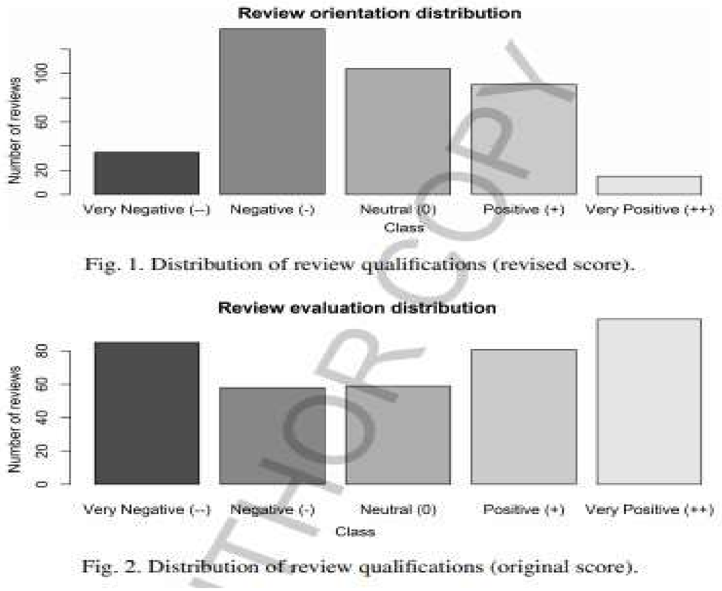


Fig. 1. Distribution of review qualifications (revised score).

Fig. 2. Distribution of review qualifications (original score).

Fig 2.2.1 Distribution of review classification(original and revised)

In this research paper the authors have done sentiment analysis on rule based techniques and machine learning algorithms. When both of these were compared then it was observed that machine learning techniques are better in terms of accuracy and other parameters.

The procedure involved selection of the data set from internet. Further rule based mechanism involved the use of sentiment vader and sentiword net. In machine learning algorithms LDA Analysis is also done on naïve bayes to improve efficiency.

Data processing and extraction on product features is also done. This is followed by extraction of reviews and building a vocabulary. After application of these techniques the performance is evaluated and decisions are taken. The accuracy from rule based mechanisms was 54.7 percent which was very low. On the other hand Naïve bayes algorithm has an accuracy of 75.2 percent which is high.

So, according to SmijaDasthis[1] study shows that machine learning algorithms are more efficient than rule based mechanisms in terms of performance. This is the result which can be concluded by this research paper.

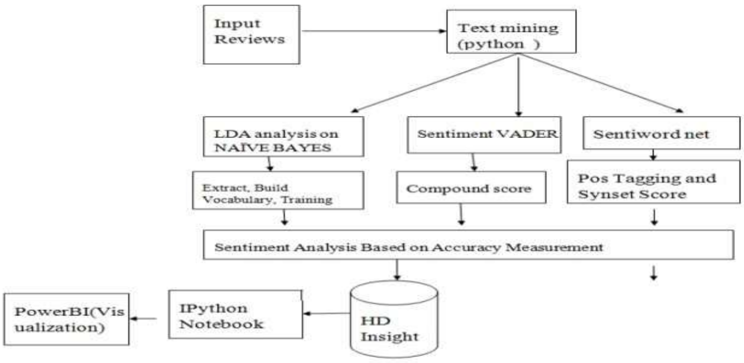
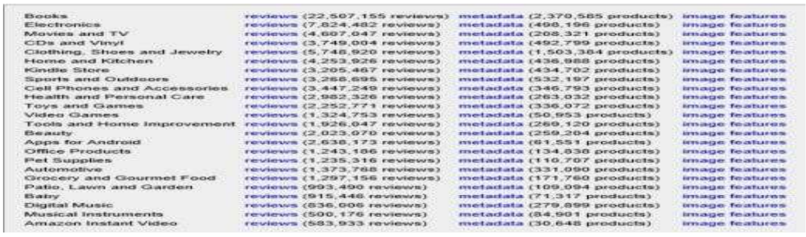

Input Reviews
Text mining (python)
LDA analysis on NAÏVE BAYES
Sentiment VADER
Sentiword net
Extract, Build Vocabulary, Training
Compound score
Pos Tagging and Synset Score
Sentiment Analysis Based on Accuracy Measurement
PowerBI(Visualization)
IPython Notebook
HD Insight

Figure 2.2.2 flowchart



Books	reviews (22,507,155 reviews)	metadata (2,370,585 products)	image features
Electronics	reviews (7,824,482 reviews)	metadata (498,196 products)	image features
Movies and TV	reviews (4,607,047 reviews)	metadata (208,321 products)	image features
CDs and Vinyl	reviews (3,749,004 reviews)	metadata (492,799 products)	image features
Clothing, Shoes and Jewelry	reviews (5,748,920 reviews)	metadata (1,503,384 products)	image features
Home and Kitchen	reviews (4,253,926 reviews)	metadata (436,988 products)	image features
Kindle Store	reviews (3,205,467 reviews)	metadata (434,702 products)	image features
Sports and Outdoors	reviews (3,268,695 reviews)	metadata (532,197 products)	image features
Cell Phones and Accessories	reviews (3,447,249 reviews)	metadata (346,793 products)	image features
Health and Personal Care	reviews (2,982,326 reviews)	metadata (263,032 products)	image features
Toys and Games	reviews (2,252,771 reviews)	metadata (336,072 products)	image features
Video Games	reviews (1,324,753 reviews)	metadata (50,953 products)	image features
Tools and Home Improvement	reviews (1,926,047 reviews)	metadata (269,120 products)	image features
Beauty	reviews (2,023,070 reviews)	metadata (259,204 products)	image features
Apps for Android	reviews (2,638,173 reviews)	metadata (61,551 products)	image features
Office Products	reviews (1,243,186 reviews)	metadata (134,838 products)	image features
Pet Supplies	reviews (1,235,316 reviews)	metadata (110,707 products)	image features
Automotive	reviews (1,373,768 reviews)	metadata (331,090 products)	image features
Grocery and Gourmet Food	reviews (1,297,156 reviews)	metadata (171,760 products)	image features
Patio, Lawn and Garden	reviews (993,490 reviews)	metadata (109,094 products)	image features
Baby	reviews (915,446 reviews)	metadata (71,317 products)	image features
Digital Music	reviews (836,006 reviews)	metadata (279,899 products)	image features
Musical Instruments	reviews (500,176 reviews)	metadata (84,901 products)	image features
Amazon Instant Video	reviews (583,933 reviews)	metadata (30,648 products)	image features

Figure 2.2.3 Dataset

Sentiment Analysis of Product-Based Reviews Using Machine Learning Approaches research paper aims to classify product reviews of various websites as negative and negative. Here the data is taken from the Amazon website.

The first step is data collection. Parts of speech are also very important in the sentence structure. Identification of negative phrases is also done. Machine learning algorithms such as Naive Bayesian classifier, Random forest, support vector, logistic regression is applied to the data used.

This research paper also considers ROC curve, Recall and precision value. According to [2]ANUSUYA DHARA and ARKADEB SAHA The paper threw light on a typical sentiment analysis model that contained three core steps, that were preparation of data, analysis of reviews and classification of sentiment . This research paper deals with the classification of texts on the basis of emotions where they are accurate.

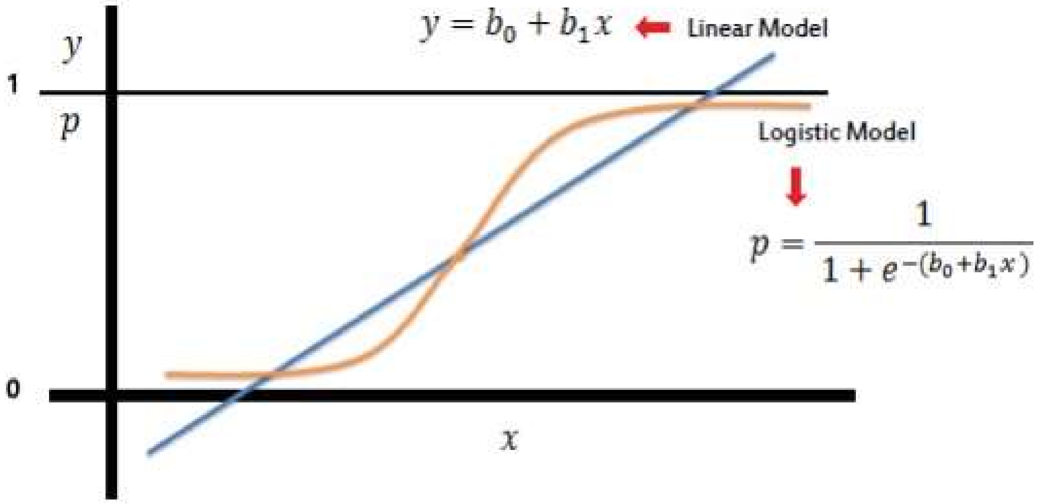


Figure 2.2.4 graph of linear and logistic model

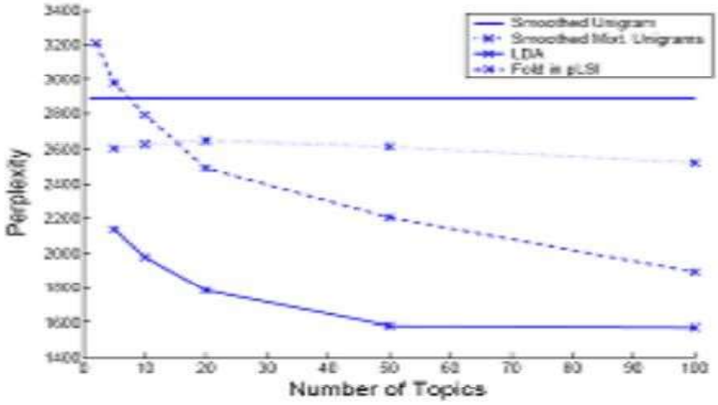


Figure 2.2.5 lda analysis

In A Review on Sentiment Analysis Techniques and Applications research paper the natural language analysis and machine learning algorithms were used for analysis of emotions.

Natural language analysis is used for text mining, interpreting and questioning. The NLP uses the speech marking section (POS), text frequency, dictionary and the weight of their assignments. It has accuracy is 80% in most cases.

Machine learning methods are more efficient than the NLP method. In the machine learning models trained in databases to analyze data and perform continuous classification, some popular machine learning algorithms are Support Vector Machine, Naive Bayes, K Nearest Neighbor etc.

2.3 Summary:

In this research papers sentiment mining is done to analyze the sentiments of a customer. These examinations of reviews are done to cater to any client needs where reviews play importance. The research paper uses machine learning techniques along with LDA Analysis on naive bayes.

Unsupervised machine learning technique is used for this purpose. This machine learning algorithm is used to categorize the positive and negative reviews. Sentiment Vader (VALENCE AWARE DICTIONARY AND SENTIMENT REASONER) is sentiment analysis instrument that is used to analyze social media data. Machine learning algorithms like Naive Bayesian classifier, Random forest, support vector, logistic regression are applied to the processed data. The research papers also takes into the consideration the ROC curve, Recall and precision value.

Natural language processing is used for text mining, making interpretations and inquiry about the data present in the dataset. NLP uses part of speeches (POS) tagging, frequency of documents, dictionary and their assigned weight. It has 80% accuracy in most of the cases.

Another method is machine learning approach which is quite efficient than NLP approach. In machine learning models are trained on the dataset to analyze the data and do further classification, some popular machine learning algorithms are Support Vector Machine, Naive Bayes,K Nearest Neighbor etc.

The techniques used in the process can be Bayesian Classifier (NB), Support Vector Machine (SVM) which comes under supervised methods and Scoring algorithms under unsupervised methods.

Scoring algorithms include binary classification, ternary classification, 5-point scale classification and aspect evaluator algorithm in which binary classification proves to better.To improve the performance and expected results longitudinal evaluation between the review and reviewer's acceptance can be proposed. After determining the consistency between the system and reviewer, sentiment analysis can be proved perfect for scientific paper reviews.

Table 2.3.1 Comparative study

TITLE	**YEAR**	**APPROACH USED**	**STRENGTH**	**GAPS IN WORK**
Tweets Classification and Sentiment Analysis for Personalized Tweets Recommendation	**2020**	XML Parser, Spell check and Tokenizer to split words, POS tagger, dependency parser.	Suitable for large datasets, provide personalized recommendations.	Overlapping of tweets**.**
Sentiment Analysis of Movie Reviews Using Machine Learning Techniques	**2017**	KNN, Random Forest and Naïve Bayes algorithm.	Comparison of the accuracies of the three algorithms, minimum error.	Small dataset
Sentiment Analysis of Customer Product Reviews Using Machine Learning	**2017**	Naïve Bayes algorithm, Support Vector Machine algorithm and Decision Tree algorithm.	Thorough comparison of the supervised algorithms	Multi class classification of reviews is not done in this research paper.
Sentiment Analysis on textual Reviews	**2018**	Sentiment vader and sentiword net and LDA analysis.	LDA analysis provides efficiency.	Less optimized results.
Sentiment Analysis for Online Product Reviews using NLP Techniques and	**2016**	K means clustering, LDMA and other machine learning	High accuracy.	Only machine learning is used.

Statistical Methods		algorithms.		
K means clustering, LDMA and other machine learning algorithms.	**2019**	Naïve Bayes and Support Vector Machine.	Better accuracy for high and complex data.	The F1 score computed in this paper is 0.64 which is quite less.
A Review on Sentiment Analysis Techniques and Applications	**2016**	OpenCvand Support Vector Machine	96% of Fmeasure.	Less accuracy as compared to machine learning applications.
Deep Learning Based Text Classification: A Comprehensive Review	**2020**	Supervised Learning models	Memory Efficient models, Common sense Models.	Complex datasets are more challenging.
. Deep Learning Approach for Sentiment Analysis of Short Texts	**2017**	ConcLstm, CNN, RNN.	Long term functionality, Short texts are classified using long short-term memory.	RNN is inefficient for many NLP processes e.g., prediction of next words.
Sentiment Analysis of Product-Based Reviews Using Machine Learning Approaches.	**2014**	Naïve bayes Classifier, Random Forest, Support Vector machine	Classification of data in three steps.	Complex datasets are more challenging, Polarity Categorization problem.

CHAPTER 3

SYSTEM DESIGN AND METHODOLOGY

3.1. System Design:

System design is the process of defining system elements such as structure, structure, components and system data based on defined needs. It is the process developing and making systems that meet the specific needs and requirements of our project.

Some of the design methods are-

1.) Architectural design:

Architectural design in this project basically emphasizes on the views and models. It also tells us about the structure of the project.

2.) Logical design:

It is done to represent the flow of data. It tells us the procedures and gives us idea of inputs and outputs.

3.) Physical design:

Physical Design is defined as how users add information to the system.

System design is a very important part of any project. Before the implementation of any project we need to have a flow chart and system design of the whole project.

System Architecture helps in identifying the architecture of the system which is made. It helps us to understand the functioning associated with a project.

It gives us a clear idea about the insight of the project. In this project we have made an architectural design of the whole sentiment analysis and opinion mining of reviews.

3.1.1. System Architecture /Diagrammatical View:

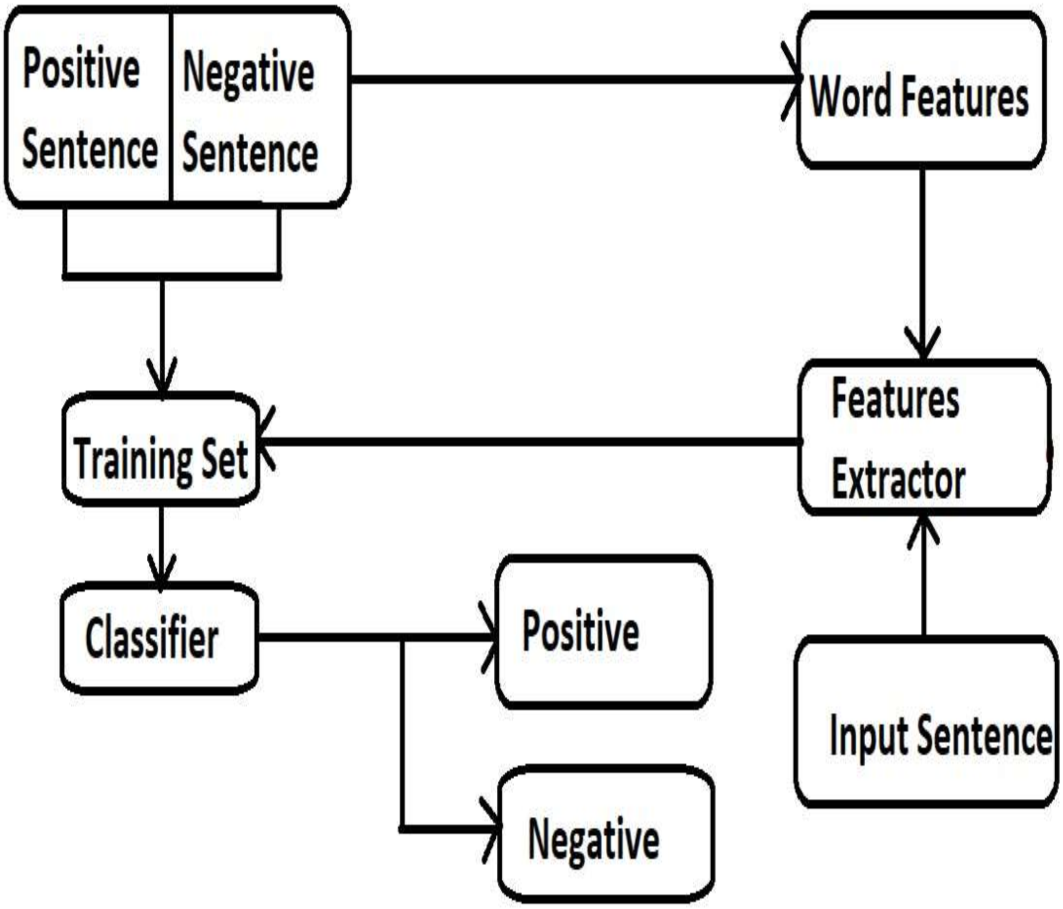


Fig 3.1.1.1 System architecture

In the above diagram, the classification of process creates a training set, from positive and negative sentences.

In the training set and input text, the text steaming and stop words are removed, now the input text and the training set can be fed into the model easily and helps the feature extractor to work accordingly and efficiently because the noise level get reduced.

After that the classifier categorizes the input text according to what it learns from the training set. Finally, the polarity of the input text is determined.

3.1.2. **Flow chart:**

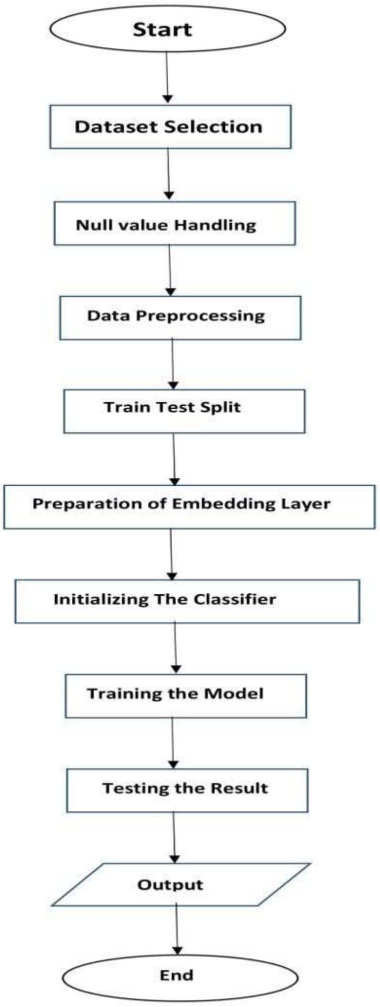


Fig 3.1.1.2 flowchart

This Flow chart depicts the outline of the project. The data to be passed in model is obtained from a product or service review website example movie review site, restaurants review site, social media platforms etc.

Public reviews with their keywords are extracted to pass in the model. Then data preprocessing is done and the data is further splitted into training data and test data.

The preparation of embedding layer is done followed by the selection of the best suited model according to accuracy and difference between training and testing accuracies.

3.2. **Algorithm:**

Neural networks are the algorithms that attempt to find the basic connections in a set of data by a process that tends to mimics the functioning of a human brain. By this sense, neural networks refer to the systems of neurons. They can either machine made or natural. Neural networks can adapt to input changes; therefore the network produces excellent results without the need to reconstruct the output processes. Neural networks are a part of artificial intelligence that has gained a lot of popularity in the recent days.

BASICS OF NEURAL NETWORK:

Neural networks are basically used to perform processing on data and it eventually tries to mimic the functioning of a human brain by creation of a model similar to that of human brain. Here our major objective is to perform computations on our data with the help of different neural network algorithms.

For this purpose we have used three different algorithms and then found out that which algorithm is best for our purpose of classifying the reviews. We first used a simple neural network and also used a embedding layer for this purpose. We compared our training and testing accuracy and found out that the difference between the two is quite more.

Gradually we changed our approach a bit and used CNN network for better accuracy and understanding of the data set but their also the difference between the training accuracy and testing accuracy was considerable and the data was over fitting.

So finally we found out that the best accuracy and less overfitting was observed when we used RNN (LSTM Algorithm). So to conclude RNN was observed as the best algorithm for sentiment analysis.

.

CHAPTER 4 IMPLEMENTATION AND RESULTS

4.1 Software and Hardware Requirements

Software Requirements:

1. **Python:** Python is an interpreter, high-level and general-purpose programming language. It was created by Guido van Rossum and it was first released in the year 1991. The language constructs and the object oriented approach helped the programmers in order to write clear and logical code for the large-scale and small-scale projects.

2. **Jupyter Notebook:** The Jupyter notebook is basically an open source web application which helps you to create as well as share the documents which contain life code visualization equations and a narrative texts. It is also used for data cleaning as well as transformation, statistical modeling, data visualization, machine learning and much more.

3. **Google Colaboratory:** Google colaboratory is basically a cloud-based service which is used for the replication of the Jupiter notebook in the cloud. It does not require any installation on the system. Google colaboratory is basically used by those readers who use something other than a desktop to work through the examples.

4. **Python Libraries:**

Matplotlib: Matplotlib is basically a library for plotting in python language. It is used in order to provide an object oriented application interface for including the plots into the applications. Most functions for plotting Matlab can easily be used in python. It includes different plots like bar plot, line plot, histogram, scatterplot etc. Using these types of plots we can easily visualize the data.

Pandas: It is basically used for the cleaning of data and its analysis. Pandas provides various features such as exploring, transforming, visualizing as well as cleaning the data. It is basically an open source python package. It is one of the most important tools for the data cleaning as well as analysis part.

Keras :Keras is a python based framework which is assumed to be the coolest python library. It is known to easily represent the neural network problems. It is a modified version

API of the tensorflow library. It is majorly used in data processing and in data visualization.

Hardware Requirements:

High CPU configuration, RAM minimum 8GB and processor 64-bit.

4.2. Implementation Details

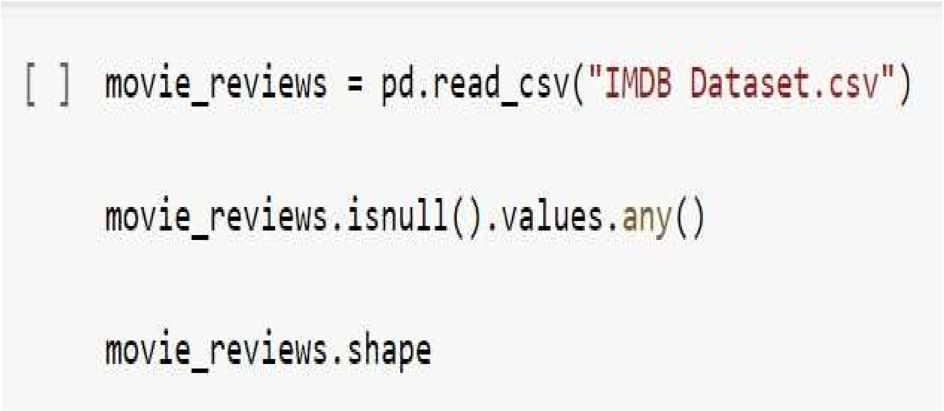
4.2.1 Snapshots of Interfaces:

1.) DATA COLLECTION:

In our project, the dataset that we have used is a CSV file which contains 50,000 records that is divided into two columns namely review and sentiment .The review column contains different reviews and the sentiment column contains sentiment for the reviews that have two values i.e. "positive" and "negative“. The dataset that we have used is downloaded from “Kaggle” website.

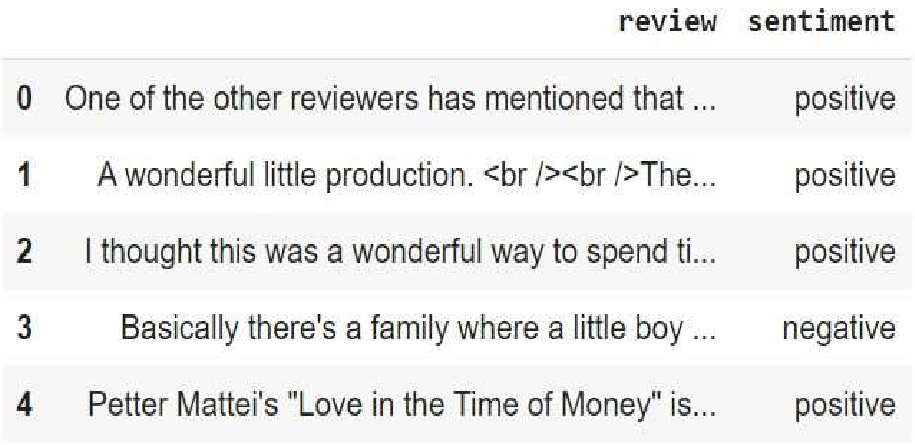
2.) IMPORTING AND ANALYZING THE DATASET:

- In the shown fig1, we have used the read_csv() method which reads our CSV file and converts it into data frame. It is taken from pandas library. Then we have checked the dataset for NULL values.
- In the shown fig2, we analyze our dataset to get an idea of what we are going to process.



```
[ ]  movie_reviews = pd.read_csv("IMDB Dataset.csv")

     movie_reviews.isnull().values.any()

     movie_reviews.shape
```

Figure 4.2.1.1 Importing dataset



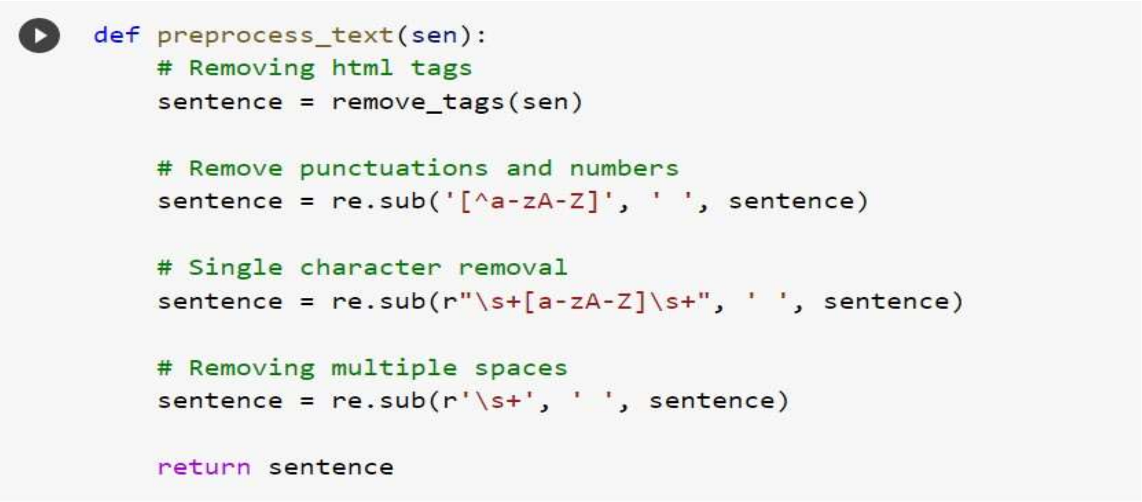
	review	sentiment
0	One of the other reviewers has mentioned that ...	positive
1	A wonderful little production. <br /><br />The...	positive
2	I thought this was a wonderful way to spend ti...	positive
3	Basically there's a family where a little boy ...	negative
4	Petter Mattei's "Love in the Time of Money" is...	positive

Figure 4.2.1.2 Snapshot of dataset

3. DATA PREPROCESSING:

- In this section, we have preprocessed our dataset in order to make the dataset consistent.

- Pre-processing is performed on the strings to remove special characters and HTML tags from the string.

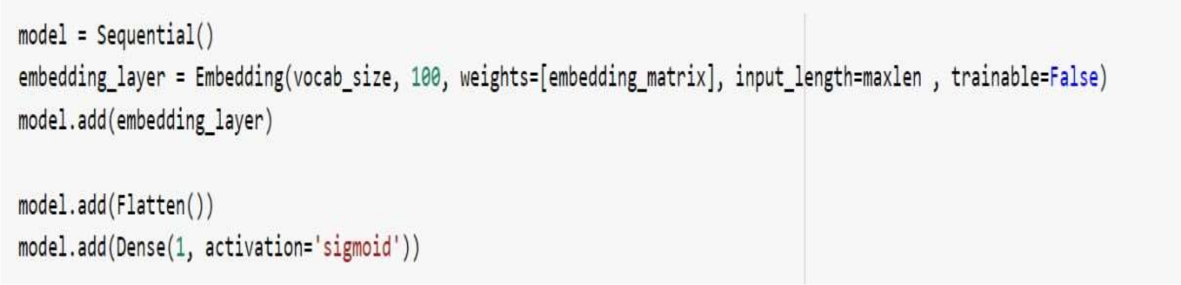


```
def preprocess_text(sen):
    # Removing html tags
    sentence = remove_tags(sen)

    # Remove punctuations and numbers
    sentence = re.sub('[^a-zA-Z]', ' ', sentence)

    # Single character removal
    sentence = re.sub(r"\s+[a-zA-Z]\s+", ' ', sentence)

    # Removing multiple spaces
    sentence = re.sub(r'\s+', ' ', sentence)

    return sentence
```

Figure 4.2.1.3 Data preprocessing

4. CLASSIFICATION WITH SIMPLE NEURAL NETWORK (SNN):

- The first deep learning model that we have developed is a simple deep neural network.
- For the implementation, we created a Sequential model and then we created our embedding layer having an input length of 100. The output vector dimension was also 100.
- Finally, we added a dense layer with sigmoid activation function.

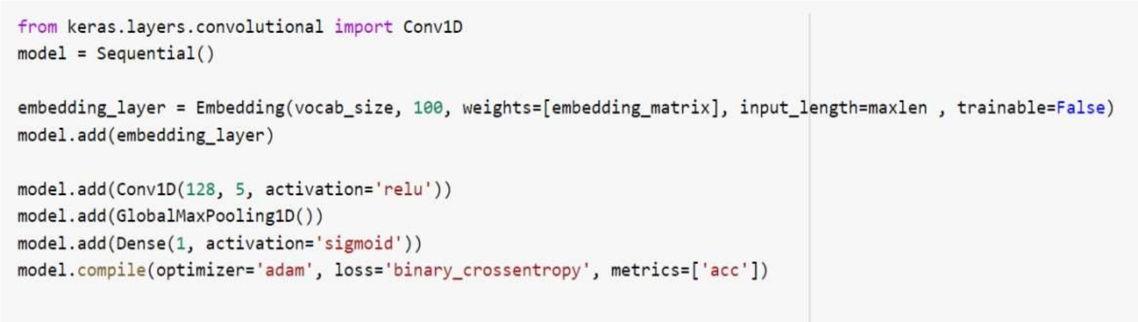


```
model = Sequential()
embedding_layer = Embedding(vocab_size, 100, weights=[embedding_matrix], input_length=maxlen , trainable=False)
model.add(embedding_layer)

model.add(Flatten())
model.add(Dense(1, activation='sigmoid'))
```

Figure 4.2.1.4 SNN model

5. CLASSIFICATION WITH CONVOLUTIONAL NEURAL NETWORK (CNN):

- The second model that we have implemented is CNN and we found that CNN works well with text data as well.
- We created a simple convolutional neural network with 1 convolutional layer and 1 pooling layer.
- Then we created a one-dimensional convolutional layer with 128 features, or kernels. The kernel size was 5 and the activation function used was sigmoid.

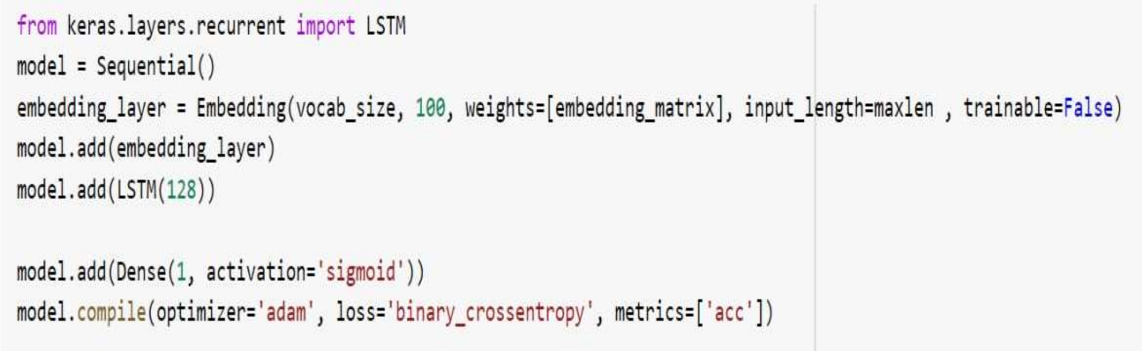


```
from keras.layers.convolutional import Conv1D
model = Sequential()

embedding_layer = Embedding(vocab_size, 100, weights=[embedding_matrix], input_length=maxlen , trainable=False)
model.add(embedding_layer)

model.add(Conv1D(128, 5, activation='relu'))
model.add(GlobalMaxPooling1D())
model.add(Dense(1, activation='sigmoid'))
model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['acc'])
```

Figure 4.2.1.5 CNN model

6. CLASSIFICATION WITH RECURRENT NEURAL NETWORK (RNN):

- Recurrent neural network is a type of neural networks that is proven to work well with sequence data.
- In this model, we have used an LSTM (Long Short Term Memory network) which is a variant of RNN, to solve sentiment classification problem.
- We have done the implementation by first initializing a sequential model and then creating an embedding layer. Next, we have created an LSTM layer with 128 neurons.

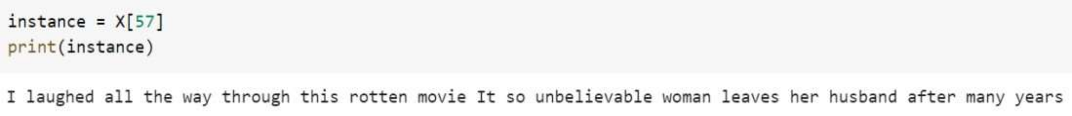


```
from keras.layers.recurrent import LSTM
model = Sequential()
embedding_layer = Embedding(vocab_size, 100, weights=[embedding_matrix], input_length=maxlen , trainable=False)
model.add(embedding_layer)
model.add(LSTM(128))

model.add(Dense(1, activation='sigmoid'))
model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['acc'])
```

Figure 4.2.1.6 RNN model

4.2.2 Test Case :

- In this section, we will see how to make predictions on a single instance or single sentiment.
- We take an example of a negative labeled review and test that review on the model and see how the model is predicting.

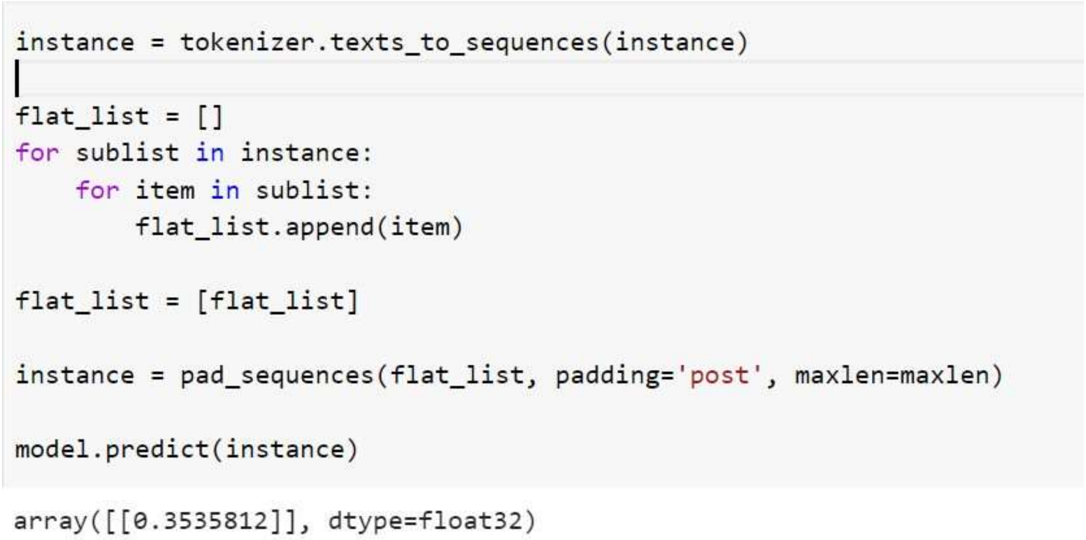


```
instance = X[57]
print(instance)
```

```
I laughed all the way through this rotten movie It so unbelievable woman leaves her husband after many years
```

Figure 4.2.1.7 Negative Review Test case

- Lets, predict the sentiment of this review.



```
instance = tokenizer.texts_to_sequences(instance)

flat_list = []
for sublist in instance:
    for item in sublist:
        flat_list.append(item)

flat_list = [flat_list]

instance = pad_sequences(flat_list, padding='post', maxlen=maxlen)

model.predict(instance)
```

```
array([[0.3535812]], dtype=float32)
```

Figure 4.2.1.8 Snapshot of the Predicted Output

- Here we have mapped the positive values as 1 and negative values as 0. For sigmoid function the floating value is in between 0 and 1 so, according to this if the value is observed to be less than 0.5 then the sentiment will be negative and if the observed value is greater than 0.5 then we can conclude that the value is positive.
- As we see here that if in some case we see that the predicted value is 0.33 that means the review is predicted as negative and actually also the review is negative only.

- Hence, we can conclude that we used three different types of neural networks to classify public sentiment about different movies. The result predicted by RNN model is better than both CNN and simple neural network.

4.2.3 Result:

1. **Comparative Study:**
    - The comparison is done between the training accuracy and test accuracy. The model performing correspondingly in training and test sets is preferable. In other words, the overfitting, which occurs when test performance is less than training performance, should be minimum.
    - We have compared three neural network modelsi.e SNN,CNN and RNN.

- In the implementation of SNN model, the flattening and embedding layers are used andwe get the training accuracy as per the implementation as 85.52% and test accuracy as 74.62% as shown in figure 4.2.1.9.

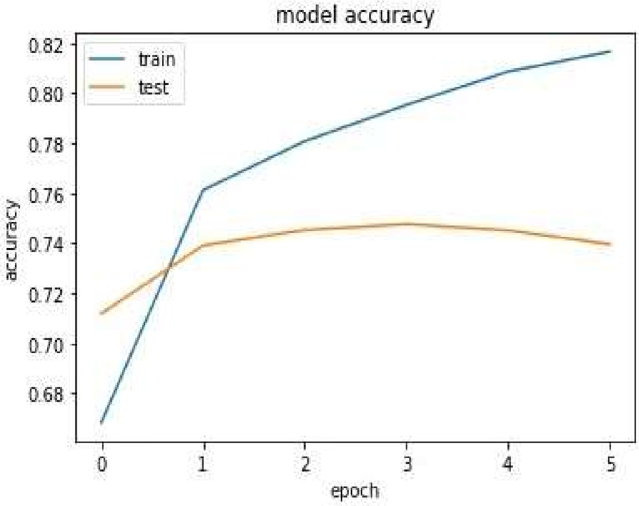
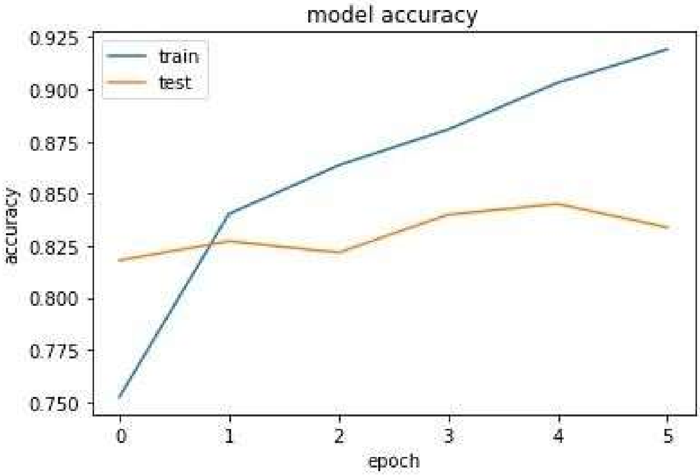


Figure 4.2.1.9 Model accuracy of SNN

- In the implementation of CNN model, no flattening layer has to be used. We get the training accuracy as 92% , test accuracy as 84.5% which are comparatively better than SNN model as shown in figure 4.2.1.10.



.

Figure 4.2.1.10 Model accuracy of CNN

- In the implementation of RNN model,We get the training accuracy as 85.40% and test accuracy as 85.09% which is comparatively better than CNN model as shown in figure 17.

- No overfitting can be seen here as difference is minimal. Hence, RNN best suits the problem.

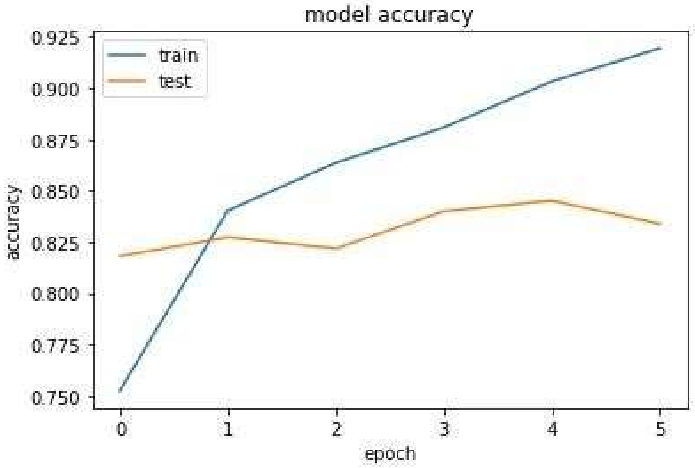


Figure 4.2.1.11 Model accuracy of RNN

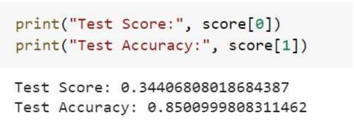
- Here is the chart showing the comparison between the three models.

Table 4.2.1.1 Comparison of training and testing accuracy

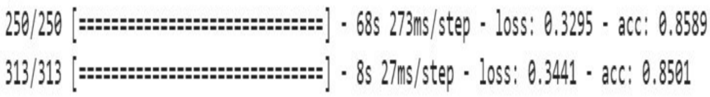
MODEL	TRAINING ACCURACY%	TESTING ACCURACY%
SNN	85.52	74.62
CNN	92	84.5
RNN	85.40	85.01

2. Accuracy:

- The best model that we get is by the RNN algorithm.
- We have used the batch size of 128 and the number of epochs used were 6.
- Once the model is trained, we get the training accuracy of 85.40% and the testing accuracy of 85.01%.
- We can see that there is a very small difference between the training accuracy and test accuracy which means that our model is not overfitting. Hence, we can conclude, that for our problem, RNN is the best algorithm.



```
print("Test Score:", score[0])
print("Test Accuracy:", score[1])

Test Score: 0.34406808018684387
Test Accuracy: 0.8500999808311462
```



```
250/250 [==============================] - 68s 273ms/step - loss: 0.3295 - acc: 0.8589
313/313 [==============================] - 8s 27ms/step - loss: 0.3441 - acc: 0.8501
```

4.1.12 Conclusion of accuracy

CHAPTER 5 Conclusion

5.1 Performance Evaluation

The project is based on opinion mining using sentiment analysis of movie reviews. Specifically, applying text classification on the reviews put on the internet, which is a part of sentiment analysis.

In the application world RNN (Recurrent Neural Network) model is basically used for sequential data or series of data. And the text classification will also provide a sequence of data. Hence RNN suits for the process.

The Performance of our project using RNN (Recurrent Neural Network) model is considered best as it results the accuracies on different data sets better than the other models like SNN (Simple Neural Network) model and CNN (Convolutional Neural Network) model. The training and testing accuracies for RNN are 85.4% and 85.09% respectively with minimum difference between them. Whereas, SNN provides training accuracy and testing accuracy as 85.52% and 74.62%. For CNN they are 92% and 84.5% correspondingly.

As discussed earlier, overfitting should be minimum i.e., training performance should be more than testing performance (with minimum difference) which also applies here. Also, the comparative analysis supports the RNN model in front of SNN (Simple Neural Network) model and CNN (Convolutional Neural Network) model.

It can be concluded that performance of RNN will support our project to produce accurate results.

5.2 Comparison with existing State-of-the-Art Technologies

In various research papers different methodologies were used for the problem of classification of reviews. It included different machine learning and deep learning methods. Various research papers that used machine learning as an approach compared the various algorithms and their accuracies and found the best approach.

They used algorithms like KNN, random forest, naive bayes, decision tree and concluded that naive bayes gives the best accuracy according to some datasets. Some projects also used LDA analysis for improving efficiency. Different approaches were used in different research papers and a comparison of all the research papers is provided in table 4.

5.3 Future Directions

Sentiment analysis and text mining is growing area of interest in the field of growing technology. Future scope is growing with the high reach of social media and internet surfing.

Using the scores calculated the future work style can be decided. Reviewing scientific papers which reflect the relation between the result and peer reviews regarding acceptance or rejection of the paper can be a area of improvement for the model. Applications having sentiment analysis as a part, predicting future events based on today's analysis are the areas.

REFERENCES:

[1] Smija Dasthis and MirsaKarim Sentiment analysis on textual reviews 2018 *IOP Conf.* IOP Publishing Ltd *Ser.: Mater. Sci. Eng.* 396 012020

[2] Anusuya dhara, SourishSengupta, Pranit Bose and Arladeb Saha Sentiment Analysis of Product-Based Reviews Using Machine Learning Approaches RCC INSTITUTE OF INFORMATION TECHNOLOGY 2017-18

[3] S.Muthukumara, A.Victoria ,Anand Mary Sentiment Analysis for Online Product Reviews using NLP Techniques and Statistical Methods International Journal of Mathematics And its Applications Volume 4, Issue 4 (2016), 303–312. ISSN: 23471557

[4] MohdRidzwan and Yaakub, A Review on Sentiment Analysis Techniques and Applications IOP Conf. Series: Materials Science and Engineering 2019 IOP Conf. Ser.: Mater. Sci. Eng. 551 012070

[5] Brian Keith Norambuena ∗ ,Exequiel Fuentes Lettura and Claudio MenesesVillegas,Sentiment analysis and opinion mining applied to scientific paper reviews, Universidad Catolica del Norte 2019

[6]AsadMasoodKhattak , RabiaBatool, Fahad Ahmed Satti, JamilHussain , Wajahat Ali Khan , AdilMehmood Khan and Bashir Hayat, Tweets Classification and Sentiment Analysis for Personalized Tweets Recommendation,2020 Hindawi Complexity Volume 2020, Article ID 8892552, 11 pages

[7] PalakBaid, Apoorva Gupta and NeelamChaplot, Sentiment Analysis of Movie Reviews using Machine Learning Techniques, 2017 International Journal of Computer Applications (0975 – 8887) Volume 179 – No.7

[8] ZeeniaSingla, SukhchandanRandhawa, Sushma Jain, Sentiment Analysis of Customer Product Reviews Using Machine Learning, 2017 International Conference on Intelligent Computing and Control (I2C2)

[9] J SaiTeja, G KiranSai, M Druva Kumar, R. Manikandan , Sentiment Analysis of Movie Reviews Using Machine Learning Algorithms - A Survey

School Of Computing, SASTRA Deemed to be University, India.

[10] KimitakaTsutsumia , Kazutaka Shimada a and Tsutomu Endo , Movie Review Classification Based on a Multiple Classifier , Department of Artificial Intelligence, Kyushu Institute of Technology, Iizuka, Fukuoka 820-8502 Japan.

[11] ShervinMinaee, NalKalchbrenner, Erik Cambria, NarjesNikzad, MeysamChenaghlu, JianfengGao.Deep Learning Based Text Classification: A Comprehensive Review, 2020 Vol. 1 No. 1.

[12]Abdalraouf, Ausif Mahmood. Deep Learning Approach for Sentiment Analysis of Short Texts.2017 3rd International Conference on Control, Automation and Robotics