PROJECT REPORT

ON

SENTIMENT ANALYSIS ON MOVIE REVIEW

Submitted by

Amit K Mahapatra – 1601209050

Saurav Panda – 1601209099

Karunesh Kumar – 1601209371

Subhendu S.Sahoo - 1601209109

7th Semester CSE / IT (Batch : 2016 – 20)

GROUP NO: CSE15

Under the guidance of

Pradipta kumar pattanayak

Silicon

…beyond teaching

DEPARTMENT OF

COMPUTER SCIENCE & ENGINEERING

SILICON INSTITUTE OF TECHNOLOGY

Silicon Hills, Patia, Bhubaneswar-751024

STUDENT DECLARATION

We hereby declare that the project report entitled **"Sentiment analysis on movie review"** submitted in partial fulfillment of the requirements for the degree of Bachelor of Technology in Computer Science and Engineering to Biju Patnaik University of Technology is our original work and not submitted to any other university or Institute for the award of any degree.

Amit K Mahapatra
Regd. No: 1601209050

Saurav Panda
Regd. No: 1601209099

Karunesh Kumar
Regd. No: 1601209371

Subhendu S.Sahoo
Regd. No: 1601209109

DEPARTMENT OF

COMPUTER SCIENCE & ENGINEERING

SILICON INSTITUTE OF TECHNOLOGY

Silicon Hills, Patia, Bhubaneswar-751024

CERTIFICATE

This is to certify that Amit K Mahapatra (Regd. No: 16012092050), Saurav Panda (Regd. No: 1601209099), Karunesh Kumar (Regd. No: 1601209371) and Subhendu S.Sahoo (Regd. No: 1601209109) have undertaken and successfully completed the project entitled **"Sentiment analysis on movies review"** under my supervision.This work is original and is being submitted as a part of 7th Semester project for the undergraduate curriculum.

Signature of the Guide:

Name :. Pradipta Kumar Pattanayak.

Designation:. Assistant Professor

DEPARTMENT OF

COMPUTER SCIENCE & ENGINEERING

SILICON INSTITUTE OF TECHNOLOGY,

Silicon Hills, Patia, Bhubaneswar-751024

ACKNOWLEDGMENT

First of all, we would like to express our thanks to our guide **Pradipta Kumar Pattanayak**, Assistant Professor, Computer Science and Engineering Department, Silicon Institute of Technology, Bhubaneswar for being an excellent mentor for us during our whole course of thesis. His encouragement and valuable advice during the entire period has made it possible for us to complete our work.

Amit Mahapatra
Regd. No: 1601209050

Saurav Panda
Regd. No: 1601209099

Karunesh Kumar
Regd. No: 1601209371

Subhendu S.Sahoo
Regd. No: 1601209109

DEPARTMENT OF

COMPUTER SCIENCE & ENGINEERING

SILICON INSTITUTE OF TECHNOLOGY

Silicon Hills, Patia, Bhubaneswar-751024

TABLE OF CONTENTS

ABSTRACT

In this project, we introduce a method to tackle the problem of sentiment classification of large movie reviews, each consisting of multiple sentences. Sentiment analysis has become important tool that can analyse review on any product or service that can be reviewed. Same goes to movie, all the audient are freely to make their own reviews on the movie that they watch and the reviews can be positive or negative based on audient satisfactions. Automated sentiment analysis is very important to make sure the analysis produce an accurate result and in faster time. By using this machine learning model, Upgrading the sentiment analysis using Neural Networks and by some modification on the number of layer with the mathematical calculation can improve the analysis accuracy. The dataset of the movie reviews will be collected on IMDB movie reviews database.

Sentiment analysis, also refers as opinion mining, is a sub machine learning task where we want to determine which is the general sentiment of a given document. Using machine learning techniques and natural language processing we can extract the subjective information of a document and try to classify it according to its polarity such as positive, neutral or negative. It is a really useful analysis since we could possibly determine the overall opinion about a selling objects, or predict stock markets for a given company like, if most people think positive about it, possibly its stock markets will increase, and so on.

In this project we choose to try to classify reviews from IMDB datasets into “positive” or “negative” sentiment by building a model based on nodes output. We implement Artificial Neural network on framework stored data for qualitative analysis purpose to get the deeper understanding of the data. Neural network uses multiple layer for reviews classification (1- positive, 0- negative) each label falls into different categories. We used Artificial neural network which gave correct results on the testing datasets with an accuracy of 85%.

Chapter 1:

Introduction

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[1].

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 analyzes 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[9].

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 centred, i.e. results of one domain cannot be applied to other 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[9].

Technically, sentiment analysis is a unique blend of artificial intelligence and machine learning, allowing organizations to use advanced tools to choose useful and reasonable moves that attract consumers toward their services and products. In order to retain customers, competitors have to track and monitor the interest of customers. Especially, not towards their own products and brands only but also towards their competitors. Machine learning have seen rapid change in previous two years with significant breakthroughs in deep learning approaches. Deep neural networks enlivened by the human brain architecture and with enough processing power these models have been shown unbelievable results on many complex problems including Natural Language Processing tasks, even without having

excessive domain knowledge. Out there many neural networks are available with their classic abilities like, Artificial Neural Network(ANN), Deep Belief Networks (DBN) with fast inferencing of the model parameters, Convolutional neural networks (CNN), and Recurrent neutral network (RNN).
In this work, we have work with Basic Artificial Neural networks(ANN) that uses special units in addition to standard units[6].

1.1.1 Need of sentiment analysis :

- **Industry Evolution:** Only the useful amount of data is required in the industry as compared to the set of complete unstructured form of the data. However, the sentiment analysis done is useful for extracting the important feature from the data that will be needed solely for the purpose of industry. Sentimental Analysis will provide a great opportunity to the industries for providing value to their gain value and audience for themselves. Any of the industries with the business to consumer will get benefit from this whether it is restaurants, entertainment, hospitality, mobile customer, retail or being travel.

- **Research Demand:** Another important reason that stands behind the growth of SA deals with the demand of research in evaluation, appraisals, opinion and their classification. Present solutions for the purpose of sentiment analysis and opinion mining are rapidly evolving, specifically by decreasing the amount of human effort that will be required to classify the comments. Also the research theme that will be based in the long established disciplines of computer science like as text mining, machine learning, natural language processing and artificial intelligence, voting advise applications, automated content analysis, etc.

- **Decision Making:** Every person who stores information on the blogs, various web applications and the web social media, social websites for getting the relevant information you need a particular method that can be used to analyze data and consequently return some of the useful results. It is going to be very difficult for company to conduct the survey that will be on

the regular basis so that there comes the need to analyze the data and locate the best of the products that will be based on user's opinions, reviews and advices. The reviews and the opinions also help the people to take important decisions helping them in research and business areas.

- **Understanding Contextual:** As human language is getting very complex day by day so it has become difficult for the machine to be able to understand human language that can be expressed in the slangs, misspelling, nuances, and the cultural variation. Thus, there will be a need of system that will make better understanding between the human and the machine language.

- **Internet Marketing:** Another important reason behind the increase in the demand of sentimental analysis is the marketing done via internet by the business and companies organization. Now they regularly monitor the opinion of the user about their brand, product, or event on blog or the social post. Thus, we see that the sentimental Analysis could also work as a tool for marketing too.

1.1.2 Applications of Sentiment Analysis :

- **Word of Mouth (WOM):** It is the process by which the information is given from one person to another person. It would essentially help the people to take the decisions. Word of Mouth has given the information about the opinions, attitudes, reactions of consumers about the related business, services and the products or even the ones that can be shared with more than one person. Therefore, this is going to be where Sentiment Analysis comes into picture. As the online review blogs, sites, social networking sites have provided the large amount of opinions, it has helped in the process of decision-making so much easier for the user.

- **Voice of Voters:** Each of the political parties usually spent a major chunk of the amount of money for the aim of

campaigning for their party or for influencing the voters. Thus if the politicians know the people opinions, reviews, suggestions, these can be done with more effect. This is how process of Sentimental analysis does not only help political parties but on the other hand help the news analysts alongside. Also the British and the American administration had already used some of the similar techniques.

- **Online Commerce:** There is vast number of websites related to ecommerce. Majority of them had the policy of getting the feedback from its users and customers. After getting information from various areas like service and quality details of the users of company users experience about features, product and any suggestions. These details and reviews have been collected by company and conversion of data into the geographical form with the updates of the recent online commerce websites who use these current techniques.

- **Voice of the Market(VOM):** Whenever a product is to be launched by a specific company, the customers would to know about the product ratings, reviews and detailed descriptions about it. Sentiment Analysis can help in analyzing marketing, advertising and for making new strategies for promoting the product. It provides the customer an opportunity to choose the best among the all.

- **Brand Reputation Management(BRM):** Sentiment analysis would help to determine how would be a company's brand,service and the service or product that would be perceived by the online community. Brand Reputation Management will be concerned about the management of the reputation of market. It has focuses on the company and product rather than customer. Thus the opportunities were created for the purpose of managing and strengthening the brand reputation of the organizations.

- **Government:** Sentiment Analysis has helped the administration for the purpose of providing various services to the public. Fair results have to be generated for analyzing the negative and positive points of government. Thus sentiment analysis is helpful in many fields like decision making policies, recruitments, taxation and evaluating social strategies. Some of

the similar techniques that provide the citizen oriented government model where the services and the priorities should be provided as per the citizens. One of the interesting problems which can be taken up is applying this method in the multi-lingual country like the India where content of the generating mixture of the different languages (e.g. Bengali English) is a very common practice.

1.2 Problem Statement

Sentiment Analysis is a process of extracting feature from user's thoughts, views, feelings and opinions which they post on any social network websites. The result of sentiment analysis is classification of natural language text into classes such as positive, negative and neutral. The amount of data generated from social network sites is huge; this data is unstructured and cannot give any meaningful information until it is preprocessed and analyzed. Thus, to make this huge amount of data useful we perform sentiment analysis, i.e. extracting feature from this data and classify them. Sentiment analysis is very necessary in today's world, as people always get affected by the thinking and opinions other people. 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. So there is a need of system that can automatically generate sentiment analysis from this huge amount of data. We are focusing on sentiment analysis on movie reviews from imdb. We are taking the dataset from tensorflow.keras.datasets.imdb which is under the tensorflow framework. We are using Artificial Neural Network with 4 layers for binary classification of the reviews by using back-propagation algorithm[8].

1.3 Motivation

In today's world where computer science and technology thrives to make lives of people relatively simpler, data science plays a huge role in helping to achieve that goal. With this project, we try to delve into the fascinating world of data science. Sentiment analysis is an intriguing, yet complex application of data science. The applications of sentiment

analysis are massive. Social media provides a better environment for sharing user experience in an interactive and informal way so most of the people show interest to post content about what they really feel. Some interesting methods are used to collect data from social media sites like Twitter, Facebook. Collecting data from these conversations is authentic for analysis purpose because the majority of the posts by people will be instinctive. Pure manual analysis and automatic analysis cannot deal with the growing scale of data. Thus the need for qualitative analysis is increasing.

1.4 Organization of Reports:

We have already discussed about the basic theoretical concepts of sentiment analysis, and how we IMDB datasets for our data in Chapter 1. We also stated the problem statement and motivation to do the project. The next section, Chapter 2 contains a review of related works by different researchers to highlight how sentiment analysis has evolved through the years. The section following that, Chapter 3 talks about Algorithm and Implementation in a detailed manner. First, we discuss about how raw data is extracted and preprocessed using several steps. Then, we talk about the classification technique, Artificial Neural Network and how it is used to train the data to predict the sentiment. We then proceed with Chapter 4 that lists the conclusion and results along with a few snapshots of our work. We also discuss future prospects of our work in this section. Finally in Chapter 5, we list the references and in Chapter 6, we list the appendix.

Chapter 2

Review of Related Work

Many research have 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.

P. Pang, L. Lee et al [2]

By collecting large amount of data has always been a key to find out what people is thinking or expecting. With the emergence in the field of social media, availability of data which is full of opinion resources is very high. Other resources such as blogs, review sites, messages, etc. are helping us to know what people can do and their opinion about the topic. The sudden increase of work in the field of data mining and sentiment extraction deals with the computational power to solve the problem of opinion mining or subjectivity in text. Hence various new systems are created based on different languages and commands that can deal directly with opinion mining as the first class object and direct response or live research also becoming the area of interest. They take a survey which covers that methodology and approaches that are used in direct response of opinion mining are more helpful than others. Their focus is on functions that can solve new challenges rising in sentiment analysis applications. They also compared these new techniques to already present traditional analysis which is based on facts.

P. Pang, L. Lee, S. Vaithyanathan et al [1]

They were the first to work on sentiment analysis. Their main aim was to classify text by overall sentiment, not just by topic e.g., classifying movie review either positive or negative. They apply machine learning algorithm on movie review database which results that these algorithms outperform human produced algorithms. The machine learning algorithms they use are NaïveBayes, maximum entropy, and

support vector machines. They also conclude by examining various factors that classification of sentiment is very challenging. They show supervised machine learning algorithms are the base for sentiment analysis.

E. Loper, S. Bird et al [3]

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.

O. Almatrafi, S. Parack, B. Chavan et al [4]

They are the researchers who proposed a system based on location. According to them, Sentiment Analysis is carried out by Natural Language Processing (NLP) and machine learning algorithms to extract a sentiment from a text unit which is from a particular location. They study various applications of location based sentiment analysis by using a data source in which data can be extracted from different locations easily. In Twitter, there is field of tweet location which can easily be accessed by a script and hence data from particular location can be collected for identifying trends and patterns.

Thomos mikolov, GregCorrado, JeffreyDean[8]

In this paper two novel model architectures for computing continuous vector representations of words from very large datasets. The quality of the sere presentations is measured in a word similarity task, and the results are compared to the previously best performing techniques based on different types of neural networks. We observe large improvements in accuracy at much lower computational cost, i.e. it takes less than a day to learn high quality word vectors from a 1.6 billion words data set. Furthermore, we show that these vectors provide state-of-the-art performance on our test set for measuring syntactic and semantic word similarities.

Chapter 3:

3.1 DATASET USED

IMDB (Internet Movie Database): It is an online database of information related to films, television programs, home videos, video games, and streaming content online including cast, production crew and personal biographies, plot summaries, trivia, fan and critical reviews, and ratings. An additional fan feature, message boards, was abandoned in February 2017. Originally a fan-operated website, the database is owned and operated by IMDb.com[4].

The dataset used for this task was collected from Large Movie Review Dataset which was used by the AI department of Stanford University for the associated publication. The dataset contains 50,000 training examples collected from IMDb[4] where each review is labelled with the rating of the movie on scale of 1-10. As sentiments are usually bipolar like good/bad or happy/sad or like/dislike, we categorized these ratings as either 1 (like) or 0 (dislike) based on the ratings. Initially the dataset was containing 50,000 reviews out of those first 25000 reviews of training , we have kept first 10000 for validation and rest 15000 for training. After the training and validation is carried out we perform the testing on next 25000 reviews and get the testing accuracy. Then we pick a single review or at times multiple reviews from test data and compare the predicted and actual value of either positive(1) or negative(0).

3.2 DATA EXTRACTION AND DATA PRE-PROCESSING:

To gather the data many options are possible. In some previous paper researches, they built a program to collect automatically a corpus of reviews based on two classes, “positive” and “negative”, by querying Social sites with two type of emoticons:

- Happy emoticons, such as “:)”, “:P”, “:-)” etc.
- Sad emoticons, such as “:(“, “:’(”, “=(“.

Others make their own dataset of reviews my collecting and annotating them manually which very long and fastidious[5]. Additionally, to find a

way of getting a corpus of reviews, we need to take of having a balanced data set, meaning we should have an equal number of positive and negative texts, but it needs also to be large enough. Indeed, more the data we have, more we can train our classifier and more the accuracy will be. After many researches, we found a dataset of 1578612 reviews in english coming from two sources: Kaggle and Sentiment140. It is composed of four columns that are ItemID, Sentiment, Sentiment Source and Sentiment Text.. We are only interested by the Sentiment column corresponding to our label class taking a binary value, 0 if the review is negative, 1 if the review is positive and the Sentiment Text columns containing the tweets in a raw format[2].

Data Pre-processing:

Machine learning algorithms take numbers as inputs. This means that we will need to convert the texts into numerical vectors. There are two steps to this process[8]:

1. **Tokenization:** Divide the texts into words or smaller sub-texts, which will enable good generalization of relationship between the texts and the labels. This determines the “vocabulary” of the dataset (set of unique tokens present in the data).
2. **Vectorization:** Define a good numerical measure to characterize these texts.

The Above 2 Steps are Exemplified Below :

Texts: 'The mouse ran up the clock' and 'The mouse ran down'
Index assigned for every token:
{'clock': 5, 'ran': 3, 'up': 4, 'down': 6, 'the': 1, 'mouse': 2}.
NOTE: 'the' occurs most frequently, so the index value of 1 is assigned to it.
Some libraries reserve index 0 for unknown tokens, as is the case here.
Sequence of token indexes: 'The mouse ran up the clock' = [1, 2, 3, 4, 1, 5]

There are two options available to vectorize the token sequences:

One-hot encoding: Sequences are represented using word vectors in n- dimensional space where n = size of vocabulary. This representation works great when we are tokenizing as characters, and the vocabulary is therefore small. When we are tokenizing as words, the vocabulary will usually have tens of thousands of tokens, making the one-hot vectors very sparse and inefficient.

Example:

'The mouse ran up the clock' =

[

[0, 1, 0, 0, 0, 0, 0],

[0, 0, 1, 0, 0, 0, 0],

[0, 0, 0, 1, 0, 0, 0],

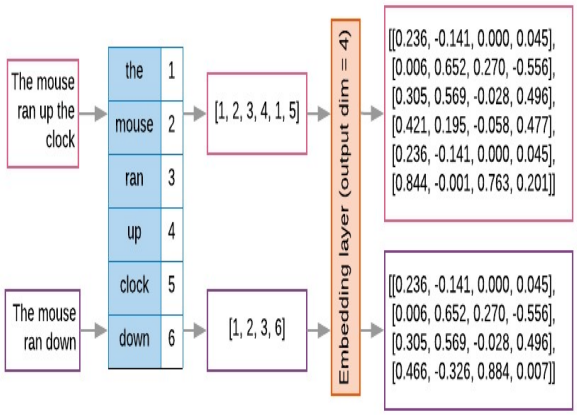
[0, 0, 0, 0, 1, 0, 0],

[0, 1, 0, 0, 0, 0, 0],

[0, 0, 0, 0, 0, 1, 0]

]

Word embeddings: Words have meaning(s) associated with them. As a result, we can represent word tokens in a dense vector space (~few hundred real numbers), where the location and distance between words indicates how similar they are semantically. This representation is called Word Embedding.



Feature selection : Not all words in the data contribute to label predictions. We can optimize the learning process by discarding rare or irrelevant words from the vocabulary. In fact, it is observed that using the most frequent 20,000 features is generally sufficient.

VI. BUILD, TRAIN, AND EVALUATE THE MODEL:

Building machine learning models with Keras(in Python) is all about assembling together layers, data-processing building blocks, much like we would assemble Lego bricks. These layers allow to specify the sequence of transformations to perform on the input. As the learning algorithm takes in a single text input and outputs a single classification, we can create a linear stack of layers using the Sequential Model API exist in Python.

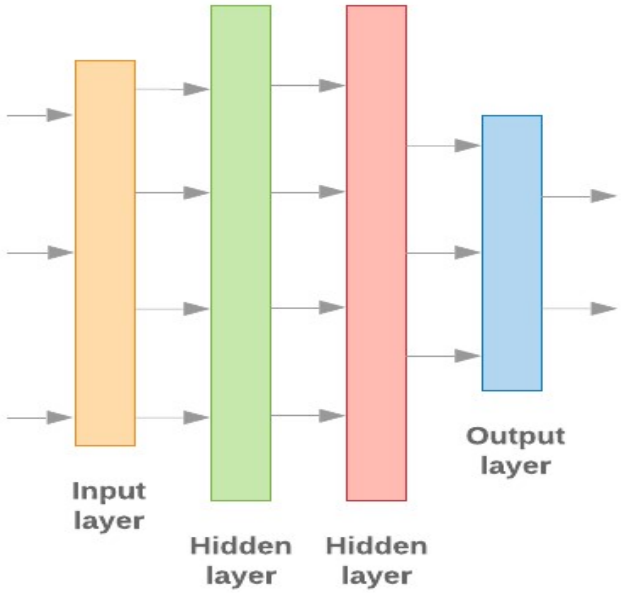


Fig 1: Basic Ann Model

Train The Model : Now that the model architecture is prepared, the model need to be trained. Training involves making a prediction based on the current state of the model, calculating how incorrect the prediction is, and updating the weights or parameters of the network to minimize this error and make the model predict better. We repeat this process until our model has converged and can no longer learn. There are three key parameters to be chosen for this process[8].

- Metric: How to measure the performance of our model using a metric.

- Loss function: A function that is used to calculate a loss value that the training process then attempts to minimize by tuning the network weights. For classification problems, cross-entropy loss works well.

Optimizer: A function that decides how the network weights will be updated based on the output of the loss function.

3.3 MODEL USED(ANN):

ANNs began as an attempt to exploit the architecture of the human brain to perform tasks that conventional algorithms had had little success. They soon reoriented towards improving empirical results, mostly abandoning attempts to remain true to their biological precursors. Neurons are connected to each other in various patterns, to allow the output of some neurons to become the input of others. The network forms a directed weighted graph[9].

Neurons:

ANNs retained the biological concept of artificial neurons which receive input, combine the input with their internal state (activation) and an optional threshold using an activation function, and produce output using an output function. The initial inputs are external data, such as images and documents. The ultimate outputs accomplish the task, such as recognizing an object in an image. The important characteristic of the activation function is that it provides a smooth transition as input values change, i.e. a small change in input produces a small change in output[9].

Connections and weights:

The network consists of connections, each connection providing the output of one neuron as an input to another neuron. Each connection is assigned a weight that represents its relative importance. A given neuron can have multiple input and output connections.

$$V(i,j) = W_0*X_0 + W_1*X_1 + W_2*X_2\ldots\ldots\ldots\ldots + Wn*Xn$$

Propagation function

The propagation function computes the input to a neuron from the outputs of its predecessor neurons and their connections as a weighted sum. A bias term can be added to the result of the propagation[13].

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

where, x is the input to the node.

Organization:

The neurons are typically organized into multiple layers, especially in deep learning, Neurons of one layer connect only to neurons of the immediately preceding and immediately following layers. The layer that receives external data is the input layer. The layer that produces the ultimate result is the output layer. In between them are zero or more hidden layers. Single layer and unlayered networks are also used. Between two layers, multiple connection patterns are possible. They can be fully connected, with every neuron in one layer connecting to every neuron in the next layer.

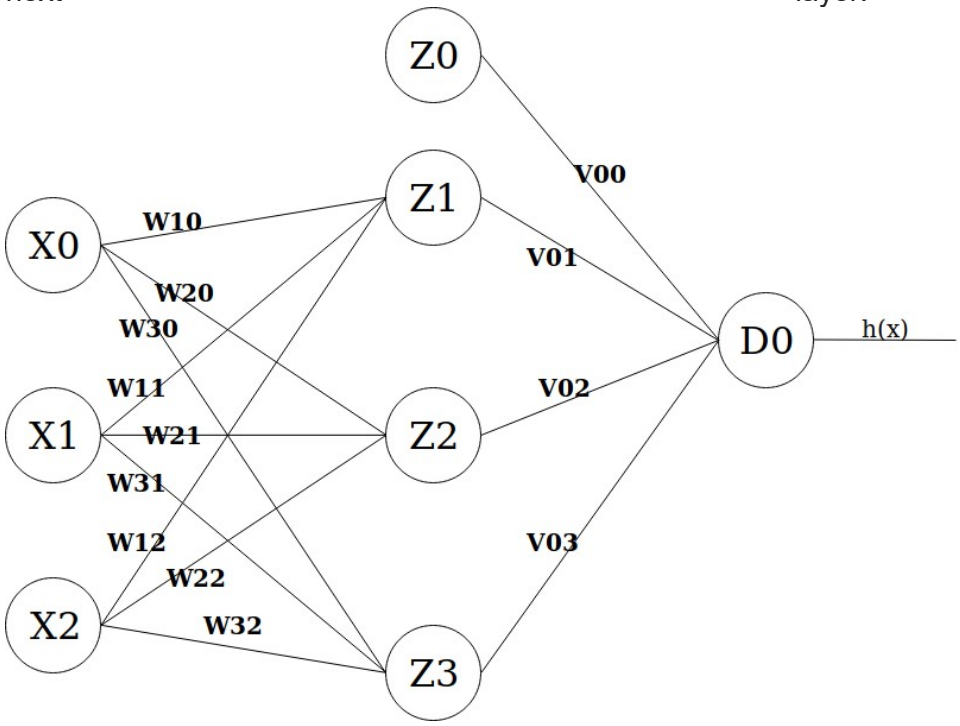


Fig 2: Artificial neural network

Learning:

Learning is the adaptation of the network to better handle a task by considering sample observations. Learning involves adjusting the weights (and optional thresholds) of the network to improve the accuracy of the result. This is done by minimizing the observed errors. Learning is complete when examining additional observations does not usefully reduce the error rate. Even after learning, the error rate typically does not reach 0. If after learning, the error rates too high, the network typically must be redesigned. Practically this is done by defining a cost function that is evaluated periodically during learning. As long as its output continues to decline, learning continues. The cost is frequently defined as a static whose value can only be approximated. The outputs are actually numbers, so when the error is low, the difference between the output and the correct answer. Learning attempts to reduce the total of the differences across the observations. Most learning models can be viewed as a straightforward application of optimization theory and statistical estimation.

Learning rate

The learning rate defines the size of the corrective steps that the model takes to adjust for errors in each observation. A high learning rate shortens the training time, but with lower ultimate accuracy, while a lower learning rate takes longer, but with the potential for greater accuracy. Optimizations such as Quick Propagation are primarily aimed at speeding up error minimization, while other improvements mainly try to increase reliability. In order to avoid oscillation inside the network such as alternating connection weights, and to improve the rate of convergence, refinements use an adaptive learning that increases or decreases as appropriate. The concept of momentum allows the balance between the gradient and the previous change to be weighted such that the weight adjustment depends to some degree on the previous change. A momentum close to 0 emphasizes the gradient, while a value close to 1 emphasizes the last change.

Cost function:

While it is possible to define a cost function ad hoc, frequently the choice is determined by the functions desirable properties (such as convexity) or because it arises from the model (e.g., in a probabilistic model the model's posterior probability can be used as an inverse cost).

Backpropagation:

It is a method to adjust the connection weights to compensate for each error found during learning. The error amount is effectively divided among the connections. Technically, backprop calculates the gradient (the derivative) of the cost function associated with a given state with respect to the weights.

Chapter 4:

4.1 Snapshots:

Results and Future Scope

```
C:\Users\Dell\Desktop\zzzzz>python test.py
<START> this film was just brilliant casting location scenery story direction everyone's really suited the part they played and you could just imagine being there robert <UNK> is an amazing actor and now the sa
me being director <UNK> father came from the same scottish island as myself so i loved the fact there was a real connection with this film the witty remarks throughout the film were great it was just brilliant
so much that i bought the film as soon as it was released for <UNK> and would recommend it to everyone to watch and the fly fishing was amazing really cried at the end it was so sad and you know what they say i
f you cry at a film it must have been good and this definitely was also <UNK> to the two little boy's that played the <UNK> of norman and paul they were just brilliant children are often left out of the <UNK> l
ist i think because the stars that play them all grown up are such a big profile for the whole film but these children are amazing and should be praised for what they have done don't you think the whole story w
as so lovely because it was true and was someone's life after all that was shared with us all
```

Fig 3: Actual Review

```
the test label is
1
[   1   14   22   16   43  530  973 1622 1385   65  458 4468   66 3941
    4  173   36  256    5   25  100   43  838  112   50  670    2    9
   35  480  284    5  150    4  172  112  167    2  336  385   39    4
  172 4536 1111   17  546   38   13  447    4  192   50   16    6  147
 2025   19   14   22    4 1920 4613  469    4   22   71   87   12   16
   43  530   38   76   15   13 1247    4   22   17  515   17   12   16
  626   18    2    5   62  386   12    8  316    8  106    5    4 2223
 5244   16  480   66 3785   33    4  130   12   16   38  619    5   25
  124   51   36  135   48   25 1415   33    6   22   12  215   28   77
   52    5   14  407   16   82    2    8    4  107  117 5952   15  256
    4    2    7 3766    5  723   36   71   43  530  476   26  400  317
   46    7    4    2 1029   13  104   88    4  381   15  297   98   32
 2071   56   26  141    6  194 7486   18    4  226   22   21  134  476
   26  480    5  144   30 5535   18   51   36   28  224   92   25  104
    4  226   65   16   38 1334   88   12   16  283    5   16 4472  113
  103   32   15   16 5345   19  178   32    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    0    0    0    0    0    0    0
    0    0    0    0]
```

Fig 4: Mapping of a single review into indices

```
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, None, 16)          160000
_________________________________________________________________
global_average_pooling1d (Gl (None, 16)                0
_________________________________________________________________
dense (Dense)                (None, 16)                272
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 17
=================================================================
Total params: 160,289
Trainable params: 160,289
Non-trainable params: 0
```

Fig 5: Model Description

```
15000/15000 [==============================] - 2s 114us/sample - loss: 0.0064 - acc: 0.9949 - val_loss: 0.1052 - val_acc: 0.8684
Epoch 23/40
15000/15000 [==============================] - 2s 121us/sample - loss: 0.0063 - acc: 0.9950 - val_loss: 0.1060 - val_acc: 0.8674
Epoch 24/40
15000/15000 [==============================] - 2s 119us/sample - loss: 0.0059 - acc: 0.9950 - val_loss: 0.1067 - val_acc: 0.8693
Epoch 25/40
15000/15000 [==============================] - 2s 111us/sample - loss: 0.0057 - acc: 0.9951 - val_loss: 0.1072 - val_acc: 0.8694
Epoch 26/40
15000/15000 [==============================] - 2s 112us/sample - loss: 0.0055 - acc: 0.9953 - val_loss: 0.1079 - val_acc: 0.8679
Epoch 27/40
15000/15000 [==============================] - 2s 112us/sample - loss: 0.0054 - acc: 0.9953 - val_loss: 0.1082 - val_acc: 0.8674
Epoch 28/40
15000/15000 [==============================] - 2s 110us/sample - loss: 0.0053 - acc: 0.9954 - val_loss: 0.1090 - val_acc: 0.8675
Epoch 29/40
15000/15000 [==============================] - 2s 108us/sample - loss: 0.0052 - acc: 0.9953 - val_loss: 0.1092 - val_acc: 0.8676
Epoch 30/40
15000/15000 [==============================] - 2s 107us/sample - loss: 0.0050 - acc: 0.9955 - val_loss: 0.1097 - val_acc: 0.8676
Epoch 31/40
15000/15000 [==============================] - 2s 116us/sample - loss: 0.0049 - acc: 0.9955 - val_loss: 0.1099 - val_acc: 0.8677
Epoch 32/40
15000/15000 [==============================] - 2s 106us/sample - loss: 0.0049 - acc: 0.9955 - val_loss: 0.1103 - val_acc: 0.8675
Epoch 33/40
15000/15000 [==============================] - 2s 109us/sample - loss: 0.0048 - acc: 0.9955 - val_loss: 0.1108 - val_acc: 0.8667
Epoch 34/40
15000/15000 [==============================] - 2s 117us/sample - loss: 0.0048 - acc: 0.9955 - val_loss: 0.1109 - val_acc: 0.8663
Epoch 35/40
15000/15000 [==============================] - 2s 107us/sample - loss: 0.0047 - acc: 0.9956 - val_loss: 0.1114 - val_acc: 0.8663
Epoch 36/40
15000/15000 [==============================] - 2s 113us/sample - loss: 0.0047 - acc: 0.9956 - val_loss: 0.1116 - val_acc: 0.8656
Epoch 37/40
15000/15000 [==============================] - 2s 116us/sample - loss: 0.0047 - acc: 0.9956 - val_loss: 0.1118 - val_acc: 0.8663
Epoch 38/40
15000/15000 [==============================] - 2s 119us/sample - loss: 0.0046 - acc: 0.9956 - val_loss: 0.1122 - val_acc: 0.8656
Epoch 39/40
15000/15000 [==============================] - 2s 115us/sample - loss: 0.0045 - acc: 0.9957 - val_loss: 0.1124 - val_acc: 0.8651
Epoch 40/40
15000/15000 [==============================] - 2s 118us/sample - loss: 0.0045 - acc: 0.9957 - val_loss: 0.1126 - val_acc: 0.8649
25000/25000 [==============================] - 1s 42us/sample - loss: 0.1209 - acc: 0.8550
[0.12092581514209509, 0.85496]
```

Fig 6: Training, validation metrices

```
review is
<START> please give this one a miss br br <UNK> <UNK> and the rest of the cast rendered terrible performances the show is flat flat flat br br i don't know how michael madison could have allowed this one on his
plate he almost seemed to know this wasn't going to work out and his performance was quite <UNK> so all you madison fans give this a miss <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <
PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <P
AD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PA
D> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD
> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD>
 <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD>
prediction: is [0.]
actual is : 0
```

Fig 7: Predicted Vs Actual Output

4.2 Conclusion

Sentiment analysis is used to identifying people’s opinion, attitude and emotional states. The views of the people can be positive or negative. Commonly, parts of speech are used as feature to extract the sentiment of the text. An adjective plays a crucial role in identifying sentiment from parts of speech. Sometimes words having adjective and adverb are used together then it is difficult to identify sentiment and opinion.

To do the sentiment analysis of movie reviews, the proposed system first extracts the reviews from IMDB datasets by user. The system can also computes the frequency of each term in the data. Using neural network predicted class labels we obtain the results.

Internet is a large source of data, which makes it more attractive for sentiment analysis. We saw how training more and more data increased the accuracy of the predictor and how it can predict simple sentences accurately. We also saw that the increase in the length and complexity of a sentence tests the validity of the predictor and may not always give correct results. Neural network uses multiple layer for reviews classification (1- positive, 0- negative) each label falls into different categories. We used Artificial neural network which gave correct results on the testing datasets with an accuracy of 86%.

4.3 Scope of future work:

Some of future scopes that can be included in our research work are:

- Use of parser can be embedded into system to improve results.
- We can also increase the classification categories(like good, very good, bad, excellent, poor, very bad) by inserting more output layer, so that we can get better results.
- A web-based application can be made for our work in future.
- We can improve our system that can deal with sentences of multiple meanings.
- We can start work on multi languages like Hindi, Spanish, and Arabic to provide sentiment analysis to more local.

Chapter 5

References:

[1] P. Pang, L. Lee and S. Vaithyanathan, “Thumbs up? sentiment classification using machine learning techniques”, Proc. ACL-02 conference on Empirical methods in natural language processing, vol.10, pp. 79-86, 2002

[2] P. Pang and L. Lee, “Opinion Mining and Sentiment Analysis. Foundationand Trends in Information Retrieval”, vol. 2(1-2), pp.1-135, 2008

[3] E. Loper and S. Bird, “NLTK: the Natural Language Toolkit”, Proc. ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics ,vol. 1,pp. 63-70, 2002

[4] H. Wang, D. Can, F. Bar and S. Narayana, “A system for realtime Twitter sentiment analysis of 2012 U.S.presidential election cycle”, Proc. ACL 2012 System Demostration, pp. 115-120, 2012

[5] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. _Foundations and trends in information retrieval_, 2(12):pages 1-135, 2008.

[6] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. _Processing_, pages 16, 2009.

[7] Niek Sanders. Twitter sentiment corpus. http://www.sananalytics.com/lab/twitter-sentiment/. Sanders Analytics.

[8] T. Mikolov. Language Modeling for Speech Recognition in Czech, Masters thesis, Brno University of Technology, 2007.

[9] https://en.wikkipedia.org/wiki/Artificialneural network.

Chapter 6

Appendix

I. List of Figures