MACHINE LEARNING THEORY
PROJECT REPORT

SENTIMENT ANALYSIS FOR AMAZON PRODUCT REVIEWS

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, MOTILAL NEHRU NATIONAL INSTITUTE OF TECHNOLOGY,
ALLAHABAD

Group Members :

1. Satya Prakash Yadav 20174179
2. Ashutosh Kumar 20174087
3. Rajat Kumar 20171059

4. Bhupendra Kumar Prajapati 20174122

OBJECTIVES:-

1. ABSTRACT

Sentiment analysis is being used widely nowadays. It helps us to know the sentiment or opinion behind a given

set of texts. Sentiment analysis of amazon product reviews is focussing on a similar topic. Here, we are trying to predict the relation between the review and the rating given by different customers on different products.

We will use two approaches here.

1.) In this approach, we are trying to deal with the problem using algorithms like Logistic Regression, Decision Tree, Random Forest Classifier, Naive-Bayes method, SVM, KNN and ensemble classifier(combination of all).

2.) In the second approach, we are using deep neural networks(Long Short Term Memory).

Using both approaches, we are getting a better understanding of how sentiment analysis works in different situations.

2. INTRODUCTION

Nowadays, sentiment analysis has become a great field of interest, and there is a lot of research going on to understand the concept of sentiment in textual resources. As you can see, the internet is full of papers on sentiment analysis and also increasing day by day. We are using

sentiment analysis on amazon product reviews on the given bunch of text. Sentiment analysis is also known as opinion mining.

In this project we are studying the opinions of people computationally, their assessment, point of views and sentiment towards organization or individuals. You can find many applications of this technique in the real world. For example, the services provided by many companies always want to find the opinions of their consumers. As people nowadays love to give their opinions on various posts so we are ended with a huge amount of opinionated text and it's a difficult task to decipher these texts. As result, normal humans have difficulty finding the relevant data from this huge amount of text. Our main objective of this project is that we have to differentiate or classify the positive and negative reviews of the buyers on the material they have purchased and tried to build a supervised learning model to polarise large datasets like we have used in this project. The dataset we have used in this project is taken from the Amazon Product Reviews. We are building several supervised models based on our feature dataset. Our model consists of traditional algorithms as well as deep neural networks such as lstm. We have studied the and compared the results from these models and got a better understanding and polarised

behaviour of customers towards the products.

3. Literature Review

Table 1. Study of sentiment analysis

S.no	Studies	Algorithm / Method Used	Description
1.	Twitter part-of-speech tagging using pre-classification hidden Markov model – 2012 (Sun et al., 2012)	Hidden Markov Model	It generally classifies the tweet as positive, negative or neutral
2.	Predicting Helpfulness Ratings of Amazon Product Reviews - 2012 (Rodak et al., 2014)	SVM and Naive Bayes	It covers automata features of data and also classified based on token and syntactic analysis
3.	Survey on Product Review Sentiment Analysis with Aspect Ranking – 2013 (Patil and Mane, 2016)	SVM	Classification is done based on aspects of the product. Every aspect are provided with aspect ranking
4.	Automatically detecting and rating product aspects from textual customer reviews-2014 (Bancken et al., 2014)	Aspectator	Special algorithm for aspect-based classification
5.	Unsupervised Opinion Mining From Text Reviews Using SentiWordNet – 2014 (Soni and Patel, 2014)	Sentiword Net	The classification is done based on aspect level which finds out aggregate scores for a particular aspect (Fixed Syntactic patterns)
6.	Ontology-based sentiment analysis of twitter posts. Expert Systems with Applications - 2015 (Kontopoulos et al., 2013)	Ensemble Approach	Classification is done based on polarity
7.	Sentiment analysis using product review data. Journal of Big Data - 2015 (Fang and Zhan, 2015)	Naive Bayes	Extracting subjective content and tackling polarity categorization problem
8.	Sentiment Analysis for Movie Reviews -2015 (Goyal and Parulaker., 2015)	Random Forest	Classified by counting the number of words repeated
9.	User Bias Removal in Fine Grained Sentiment Analysis - 2016 (Wadbude et al., 2016)	SVM	Normalizing each user review score with respect to mean and standard deviation of all products rated by the user
10.	A Twitter Sentiment Analysis for Cloud Providers: A Case Study of Azure - 2016 (Qaisi and Aljarah, 2016)	Naive Bayes	AWS and Azure. The Opinion of customers around each one of them

An idea was proposed by a data scientist, Han. H. He

proposed that neural networks can be used to store contextual information which later can be used for sentiment classification. The authors improved sentence visualization using semantic details. Using three different datasets (Imdb, Yelp 2013, Yelp 2014), they gave a comparative analysis.

Majumder, another data scientist, proposed about multimodal sentiment analysis. It involved fusion strategy. Twitter rank algo was used by Sheik. Using this algorithm, influential users for an item can be known.
Semicircles were used in order to capture entity-level sentiment and then they proved that their methodology gives better accuracy than supervised methods.

4. METHODOLOGY

4.1. Naive Bayes

It is one of the standard generative learning algorithms for classification problems. According to Naive Bayes assumption, there is no conditional dependency between x's and given y.

$$p(x_1, ..., x_k|y) = \prod_{i=1}^{k} p(x_i|y)$$

It takes a list of positive integers (including zero). It calculates the conditional probability of each xi given y, where it takes the help of multinomial distribution in our first representation of review texts.
For including negative inputs, also we have used the glove dictionary, and we choose to model p(xi|y) based on Gaussian distribution rather than the previous one.

4.2. K-nearest neighbors

K-nearest neighbors is the algorithm used for regression and classification problem.K-nearest uses the concept that similar items are to be kept in one class. It uses different measures like closeness, proximity, distance to group the data sample. It doesn't use any parameters to adjust to its algorithm.

It initializes K according to class distributions in the dataset, then find the distance of each point with these K neighbours.Based on majority voting we assign data sample point to that neighbour(class label).

4.3. Support Vector Machine

Support Vector Machine is a supervised learning algorithm used for separating the dataset based on the Kernel function which is used in its classifier for different class categories. It uses the concept of hyperplane to separate the dataset into classes. Hyperplane which has a large margin from dataset points is used. Kernel function plays an important on how should be our hyperplane is. For classes greater than two we use exponential kernel function to separate more accurately.

4.4. Logistic Regression

Logistic regression is an algorithm for non-linear dependencies on input data. There exists different extensions of Logistic Regression we have used Ordinal Logistic Regression which is used for more than three categories to classify.
Mathematical Equation for Logistic regression is:-

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

here, y is our predicted output for input x, and b is the coefficient value for adjusting the graph. The output of Logistic Regression is between 0 to 1.We use decision boundaries to range between 0 to 1 for each class to classify our data sample.

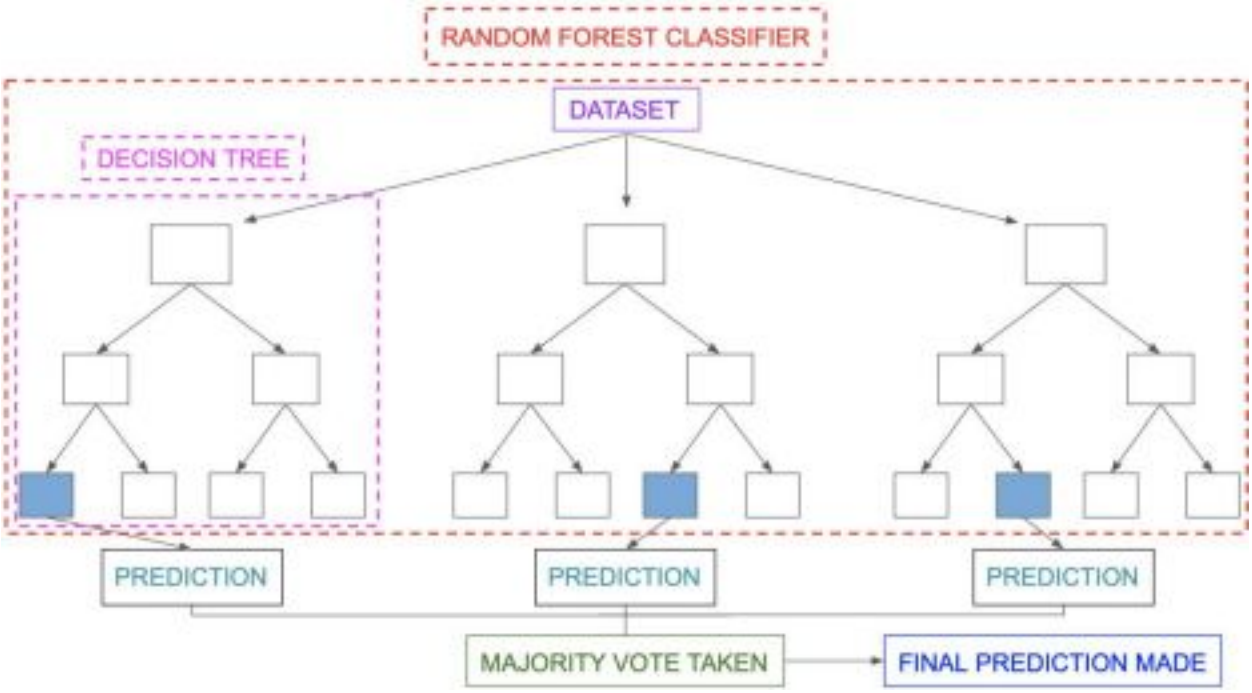
4.5.Decision Tree

Decision Tree is a very important algorithm based on a tree-based structure. It uses information gain value for each attribute to form the internal nodes of the tree. Leaf nodes are target labels of the dataset. The algorithm is a good example of decision control statements. The nodes are decided based on the information gain value, attribute which has maximum information gain value is made as selection criteria for the root. Other nodes are selected based on the same way traversing from top to bottom. For each data sample, we find it's class label from traversing through root to leaf.

4.6.Random Forest Classifier

Random forest classifier is a algorithm depends upon decision trees.In this algorithm some estimated numbers of decision trees should be constructed from arbitrary subset of given dataset.For a new data sample to classify

We find the class label on each decision tree.The most occured class label (voting) is our answer.



4.7.Ensemble Approach

Ensemble method is a approach based on majority voting and hence known as majority classifier. In this approach we use all individual classifiers results and form matrix . In this methodology we work on majority voting or mode values of different classes for each review. This method provides better accuracy than individual algorithms.

For each review we find the freq value of each class predicted by individual algorithms . We then classify the review to the class label which have maximum frequency.

It should be noted that the max votes for a particular

review should be more than 50 percent otherwise this method will not give stable prediction for that input.

STEPS:

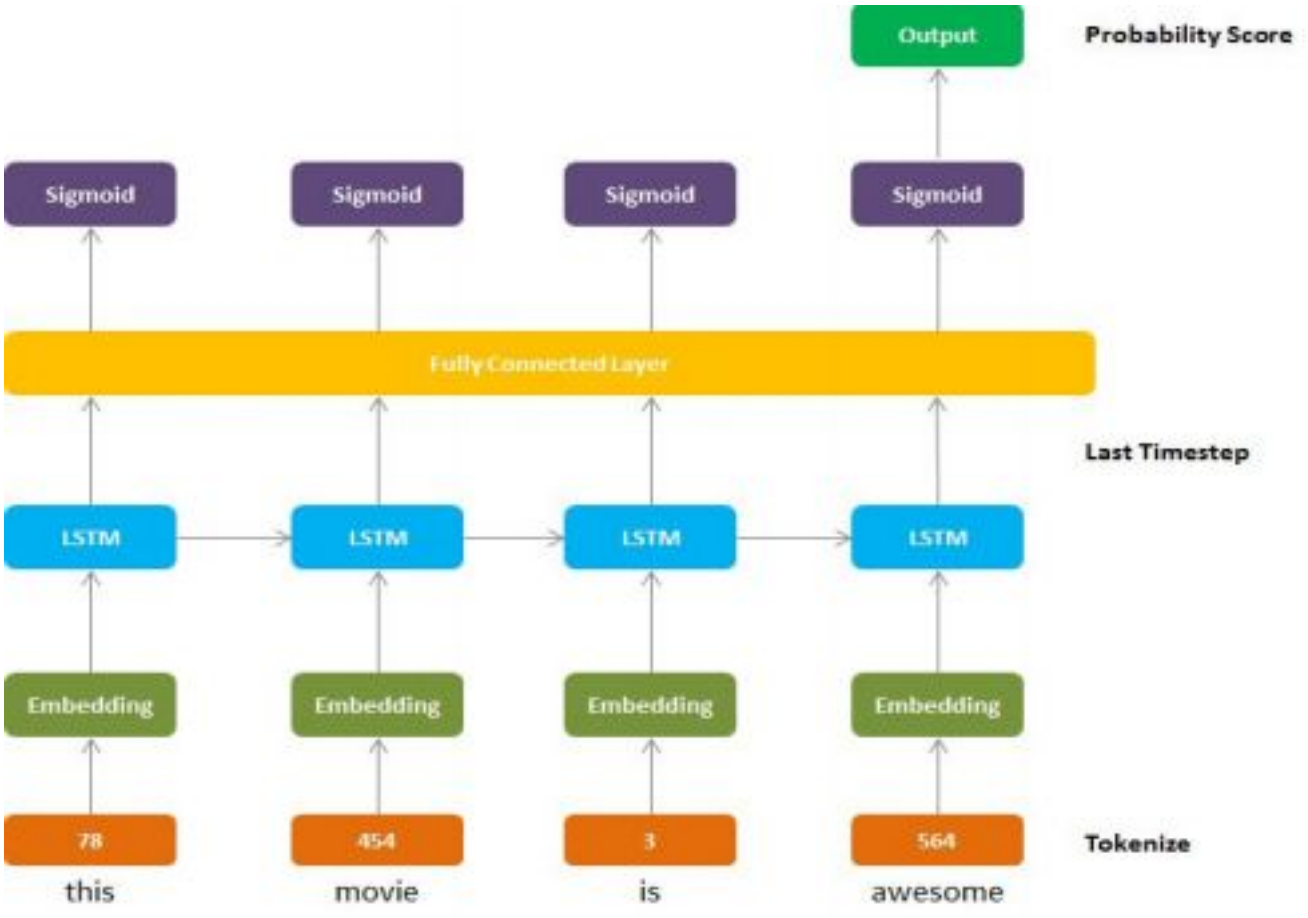
Step 1: Combine the result from all individual algorithms calculated above and form a matrix.

Step 2: Find the most frequenced class label for each sample.

Step 3: Compute Accuracy using the obtained list in step2.

4.8.Long Short Term Memory

LSTM (Long short term memory) is a more advanced unit of Recurrent neural networks which uses a series of cell states and memory gates to control flow of information. It's structure contains three gates, input , forget and output gate from which forget gate is most important.It's main concepts uses remembering of previous cell states information to predict the current state ouput.It uses activation functions to scale the data and for better control over selection of inputs based on their importance towards output.This method gives very good over accuracy time-series data , classifying problems , predictions problems , etc.



In implementation of this model , we have taken 128 hidden LSTM units and a dense layer to predict the class rating from 1-5 , which uses ‘relu’ (Rectified layer unit) as a activation function.We have trained the model for five iterations or we can iterate for more depending upon accuracy.We have used dropout rate of 0.2 to get rid of overfitting the model.

Model’s batch size is 16 , learning rate is 0.0001 so that it cannot converge frequently , and used a Adam optimizer to optimize the given set of parameters.

5.DATA SET DESCRIPTION The dataset has been obtained from Amazon Product Reviews. Number of data points in the dataset is 34660. Each datapoint has its type(name of product) and review and rating of the product. Critical portions of data like rating and review have been extracted for effective utilization.

There was noise in the data i.e. few reviews didn’t have ratings. These noises need to be eliminated. So, we are left with 34620 data points (as there were a total 40 such noise). We have plotted a graph of the ratings which shows the distribution ratio of the reviews and ratings. In the below figure, we can clearly see that there are total 5 types of ratings.

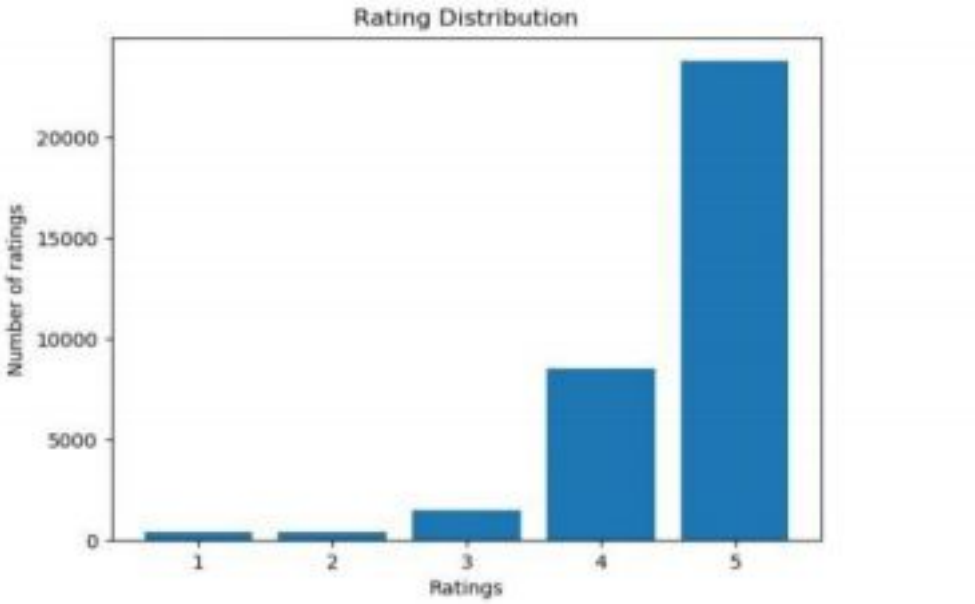


Figure 2. Rating Distribution of Amazon Reviews

There are four features in the Amazon consumer reviews dataset. These include reviews.rating as rating given by the consumer, reviews.text as feedback of product by the consumer, reviews.title as a description of the product by the company, reviews.username as the name of the product.

6.Result and Discussion

Analysis of sentiments is a field that explores emotion, opinion, or thought process about certain entities. Our paper tackles the arrangement and rearrangement of various machine learning techniques in the categorization

of sentiments. The collection of data for this project is done from the e-commerce platform Amazon after analyzing the data based on data preprocessing, pre-clarifying, and accuracy. The research work in this field is done on data format like document, sentence, and based on features of sentiment.
We have followed two strategies to compute the accuracy over the given dataset:-

First, we have applied different classifiers, which include KNN, Linear SVM, Logistic Regression, Decision tree, Random Forest, and at last using majority voting, we have implemented an ensemble approach. We have calculated the accuracy of each classifier given dataset individually and compared it with the ensemble approach. We tend to find that :

1. Using one base learner, we can produce several abstractions in combined approach.
2. More precise and efficient results have been obtained using a combined approach rather than stand-alone algorithms.
3. Overfitting can be avoided in the ensemble approach. 4. The combined method is more efficient in hiding faults of individual approaches, from which it is formed.

5. As the ensemble method cuts off biases, it reduces the deviation from the mean value.

The system we produced in our work is focused on achieving more precise results than what is already out there. We observed that accuracy is highly dependent on the count of classifiers we join for prediction. If we get more preciseness and efficiency, than it can be used for recommendation.

Second, we have approached the problem using LSTM, a unit of RNN (Recurrent Neural Networks).

As the results on the test data shows, LSTM networks are the most suitable for binary sentiment analysis on amazon product reviews. Based on the results on the evaluation datasets, we can conclude that LSTM performs very well (accuracy > 0.90) for binary classification, and that does not depend strongly on the type of product where the reviews come from. LSTM network both performs accurate results for positive and negative classes.

The conclusion from both the strategies:-
As we have seen in both strategies that both are accurate

in their results, but Recurrent neural networks with LSTM Outperforms the Ensemble algorithm. In RNN with LSTM, we have memory storage for saving previous state results(i.e., It undergoes backpropagation) for more accurate prediction, which is not present in the ensemble algorithm.

7.References:

1. Systematic literature review on context-based sentiment analysis in social multimedia

2.https://ieeexplore.ieee.org/abstract/document/8995164

3.https://dx.doi.org/10.33889/IJMEMS.2019.4.2-041