A MINI-PROJECT REPORT

ON

"Summarization & Sentiment Analysis of German Amazon Reviews"

BY

Aayush Singh
Deepanshu Sonparote
Deepali Zutshi

Under the guidance of

Internal Guide
Prof. Suresh Mestry

MANJARA CHARITABLE TRUST
RAJIV GANDHI INSTITUTE OF TECHNOLOGY
Juhu-Versova Link Road Versova, Andheri(W), Mumbai-53

Department of Computer Engineering

University of Mumbai
April- 2020

MANJARA CHARITABLE TRUST

RAJIV GANDHI INSTITUTE OF TECHNOLOGY

Juhu-Versova Link Road Versova, Andheri(W), Mumbai-53

CERTIFICATE

Department of Computer Engineering

This is to certify that

1. Aayush Singh
2. Deepanshu Sonparote
3. Deepali Zutshi

Have satisfactory completed this project entitled

"Summarization & Sentiment Analysis of German Amazon Reviews"

Towards the partial fulfilment of the

FINAL YEAR BACHELOR OF ENGINEERING
IN
(COMPUTER ENGINEERING)

as laid by University of Mumbai.

Guide **H.O.D**

Prof. Suresh Mestry **Dr. Satish Y. Ket**

Principal

Dr. Sanjay Bokade

Project Report Approval for B. E.

This project report entitled" **Summarization & Sentiment Analysis of German Amazon Reviews*"*** by ***Aayush Singh, Deepanshu Sonparote and Deepali Zutshi*** is approved for the degree of ***Bachelor of Computer Engineering***.

Examiners:

1. _____

2. _____

Date:

Place:

Declaration

We wish to state that the work embodied in this project titled ”Summarization & Sentiment Analysis of German Amazon Reviews” forms our own contribution to the work carried out under the guidance of ”Guide Name” at the Rajiv Gandhi Institute of Technology.

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

(Students Signatures)

Aayush Singh (B-858) _____

Deepanshu Sonparote (B-859) _____

Deepali Zutshi (B-870) _____

Abstract

The end of the previous decade led to a booming expansion of Amazon as a leading retail company. With the internet coverage reaching out to remote locations, the visits on such online retailers have increased exponentially. As a general tendency, people look at suggestions and feedback from the crowd to decide on whether to purchase commodities online. The paper utilizes German as its core language of analysis. This paper describes a technique to provide users with concise and accurate reviews of the product. It also provides a general outlook, which the prior customers have, of the product. The technique utilizes the classification of original and summarized reviews to training a model for review sentiment analysis using various Machine Learning models. The model, being one of a kind, utilizes data gathered by web scraping. The Encoder-Decoder LSTM model operates on this to create a summary of the reviews. The model performs abstractive summarization to retain the sense of statements. The same operations are performed on the original reviews. These two summaries are then fed to a sentiment analyzer to generate polarity of the reviews.

Contents

LIST OF FIGURES

List of Tables

CHAPTER 1

Introduction

1.1 Introduction Description

Amazon is a global giant in online retail and a prominent cloud service provider. The company that started as an online bookseller has today expanded to sell a very wide variety of consumer goods. The website allows its users to review and give feedback about the listed products and has invested significant resources to protect the integrity of these reviews. A study conducted last year found that 79% of the consumers trust online reviews as much as they would trust a personal recommendation, 85% of the consumers say that they read online reviews for local businesses and 73% of the consumers say reading a positive review about a product makes them trust the business or product more.

The customers generally write in-depth reviews about the product and on multiple occasions, potential buyers going through these reviews find it tedious to read the entire review. Due to this, the buyer might not know about all the aspects, both positive as well as negative, that they should be considering before buying the product. A solution to this problem would be the creation of a model that creates a summary of the review by analyzing the sentiments expressed by the user. It would involve detecting the polarity of the text and classifying the review as either positive or negative to identify the sentiment of the customer towards the product. The use of abstractive summarization along with sentiment analysis would make it possible to generate an entirely new summary of a long review using words that may not even appear in the text.

Doing this would allow the customer to read less data but still gain the most important information needed to decide on the product. It would also improve productivity by speeding up the surfing process of the user. This project aims to create such a system that would help the customers get a summary of reviews and assist in deciding on the product.

1.2 Organization of report

Ch.1 Introduction: This chapter introduces the subject to the reader as well as describes the solution to the problem briefly.

Ch.2 Literature Review: In this chapter, existing models are studied and a problem statement is defines for the project along with objectives and the scope.

Ch.3 Proposed System: Here, the various steps involved in the creation of the project are discussed.

Ch.4 Results & Discussion: The results generated by the software are studied and discussed n detail.

CHAPTER 2

Literature Review

2.1 Survey existing system

So far, there are a lot of research papers related to product reviews, sentiment analysis or opinion mining. For example, Xu Yun [1] el al from Stanford University applied existing supervised learning algorithms such as perceptron algorithm, naive bayes and supporting vector machine to predict a review's rating on Yelp's rating dataset. They used hold out cross validation using 70% data as the training data and 30% data as the testing data. The author used different classifiers to determine the precision and recall values. In paper [2], Maria Soledad Elli and Yi-Fan extracted sentiment from the reviews and analyze the result to build up 1 a business model. They claimed that this tool gave them pretty high accuracy. They mainly used Multinomial Naive Bayesian (MNB) and support vector machine as the main classifiers. Callen Rain [3] proposed extending the current work in the field of natural language processing. Naive Bayesian and decision list classifiers were used to classify a given review as positive or negative. Deep-learning neural networks are also popular in the area of sentiment analysis. Ronan Collobert[4] et al used a convolutional network for the semantic role labeling task with the goal avoiding excessive task-specific feature engineering. On the other hand, in paper [5], the authors proposed using recursive neural networks to achieve a better understanding compositionality in tasks such as sentiment detection.

Although a lot of research has been done on the sentiment analysis of reviews on various products, language-specific analysis of reviews is quiet uncommon. With this paper we aim to create features which would allow the user to summarize a long review of a product on Amazon and then analyse it to judge the tone/sentiment of the reviewer thus getting a general feedback about the product from the customer.

2.2 Problem Statement and Objectives

As online marketplaces have been popular during the past decades, the online sellers and merchants ask their purchasers to share their opinions about the products they have bought However, as the number of reviews available for a product grows, it is becoming more difficult for a potential consumer to make a good decision on whether to buy the product. Different opinions about the same product on one hand and ambiguous reviews on the other hand makes customers more confused to get the right decision. Here the need for analyzing this content seems crucial for all e-commerce businesses. Summarization & Sentiment Analysis of German Amazon Reviews aims to create a summary of an Amazon product review and with the help of this summary, perform sentiment analysis of the review in order to find out the outlook of buyers regarding the product, specifically for German language.

2.2.1 Objectives

- Create a summary of reviews posted on Amazon for particular product.
- Perform sentiment analysis on the reviews.
- Provide potential buyers an insight into the opinion of previous buyers.

CHAPTER 3

Proposed System

3.1 Data Collection

The dataset used for analysis is gathered manually using web scraping. By using Web Scraper extension on Google Chrome, we created a sitemap that shows how the website should be traversed and what data should be extracted. Data was collected from amazon.de official website across various products of different categories. The collected information included 3314 rows and 10 columns out of which 3 ('title'-title of the review, 'content'- actual review by customer, 'rating'- rating given to the product by customer) were used for the analysis.

	web-scraper-order	web-scraper-start-url	author	title	date	content	rating	next	next-href	Unnamed: 9
0	1582056286-2631	https://www.amazon.de/Apple-iPhone-11-128-GB-S...	MHS	Das Beste iPhone aller Zeiten	5. Januar 2020	Ich bin sehr zufrieden mit dem iPhone 11. Der ...	5,0 von 5 Sternen	Weiter→	https://www.amazon.de/Apple-iPhone-11-128-GB-S...	NaN
1	1582056184-2351	https://www.amazon.de/Apple-iPhone-11-128-GB-S...	Amazon Customer	besser als beim hersteller	21. September 2019	gestern bestellt, heute geliefert. besser geht...	5,0 von 5 Sternen	NaN	NaN	NaN

Fig 3.1: Scraped Data

3.2 Data Pre-processing

3.2.1 Cleaning

The official ratings in the dataset were of type <str> (5.0 von 5 Sternen) which were converted to integer i.e.(5,4). In addition to this, an extra column 'feedback' is added to the dataset; feedback column has values 0 or 1 assigned depending upon the ratings i.e. if rating is in range of 3 to 5 - value 1 is added to the feedback column and if rating is in range 1 to 2 - value 0 is added to the feedback column. This approach is taken because usually the rating is between 3 - 5 is considered to be good and thus value 1 indicates positive rating whereas, rating between 1 - 2 is considered to be bad and thus value 0 indicates negative rating.

	title	content	rating	feedback
0	Das Beste iPhone aller Zeiten	Ich bin sehr zufrieden mit dem iPhone 11. Der ...	5.0	1
1	besser als beim hersteller	gestern bestellt, heute geliefert. besser geht...	5.0	1
2	Gutes Handy mit kleinen Schwächen	Ich mach es mal kurz:\nGut: Optik, Verarbeitun...	4.0	1
3	Ein sehr edles Stück dieses IPHONE 11	Amazon hat wieder super-schnell geliefert. Dan...	5.0	1

Fig 3.2: Cleaned Data

3.2.2 Applying NLP Techniques for Pre-processing

1. Convert all text to lowercase
The lower() method returns the lowercase string from the given string. It converts all uppercase characters to lowercase.

2. Removing all punctuation using Python library re
Format words and remove unwanted characters.

3. Tokenization of the text using Python library spacy
Tokenization is the process of tokenizing or splitting a string, text into a list of tokens.

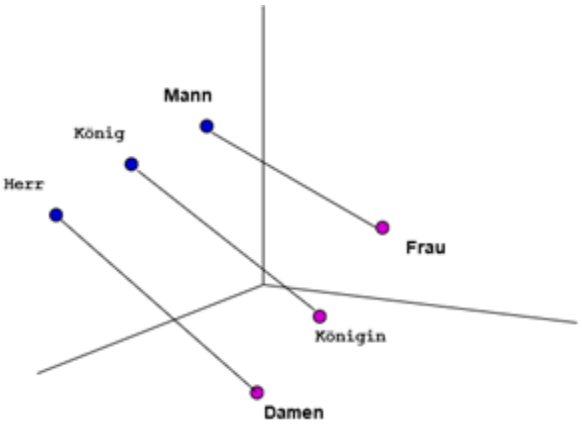
4. Removing stop words
Removing unnecessary words by using German stop words imported from spacy-de library.

3.3 Text Summarization

The reviews on Amazon can sometimes be quite lengthy and difficult to understand, to tackle this problem we used an abstractive text summarization approach. The abstraction technique entails paraphrasing and shortening parts of the source document, it creates new phrases and sentences that relay the most useful information from the original text.

3.3.1 Word Embeddings

Embedding a word involves representing it in the form of vectors. It employs mathematically embedding from space with one dimension per word to a continuous vector space in a much lower dimension. The purpose of learning word embeddings is to capture all the semantically, hierarchical and contextual pieces of information in a particular word.



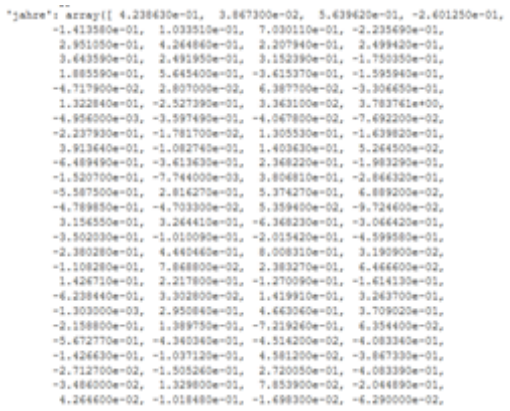


Fig 3.3: Word Embeddings

We have used GloVe pre-trained vectors on German text by [https://deepset.ai/german-word-embeddings] which gives embedding vectors for over 1309281 words. It takes a word as input and produces a vector as output, one vector per word. It helps us analyze word vectors mathematically, so these high dimensional vectors represent words and each dimension encodes a different property

like gender or type, the magnitude along the axis represents the relevance of that property to a word, similarity between words can be found using the distance between vectors. The word Mann and Frau (Man and Woman) have a similar distance to König and Königin (King and Queen) and thus can be assumed to be similar.

3.3.1.1 Vectorization

The word embeddings obtained are then used to initialize an embedding matrix which is tokenized unique vocabulary from the training data. We initialize it with zeros and copy all Glove weights of the words that are in our training dataset vocabulary. For every word outside this embedding matrix, we will find the closest word inside the matrix by measuring the cosine distance of Glove vectors.

3.3.2 Encoder-Decoder LSTM seq2seq model

The Encoder-Decoder architecture is mainly used to solve the sequence-to-sequence (Seq2Seq) problems where the input and output sequences is of different lengths. To build a text summarizer where the input is a long sequence of words (in a text body), and the output is a summary.
Our objective is to build a text summarizer where the input is a long sequence of words (in a text body), and the output is a short summary (which is a sequence as well). So, we can model this as a Many-to-Many Seq2Seq problem

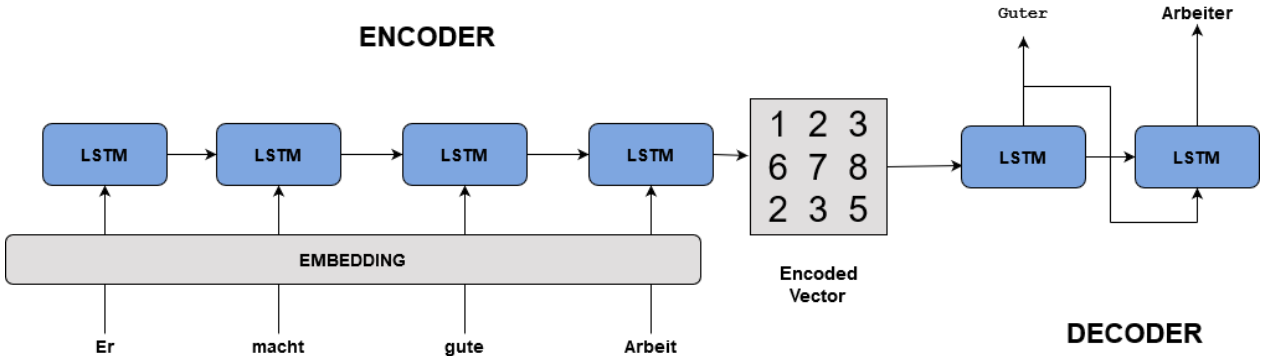


Fig 3.4: LSTM Encoder Decoder

There are two major components of a Seq2Seq model:

- Encoder
- Decoder

Long Short Term Memory (LSTM) are capable of capturing long term dependencies by overcoming the problem of vanishing gradient. We can set up the Encoder-Decoder in 2 phases:

- Training phase
- Inference phase

3.3.2.1 Training phase

In the training phase, we will first set up the encoder and decoder. We will then train the model to predict the target sequence offset by one time step. Let us see in detail on how to set up the encoder and decoder.

i. Encoder

An Encoder Long Short Term Memory model (LSTM) reads the entire input sequence wherein, at each time step, one word is fed into the encoder. It then processes the information at every time step and captures the contextual information present in the input sequence.

The below diagram which illustrates this process:

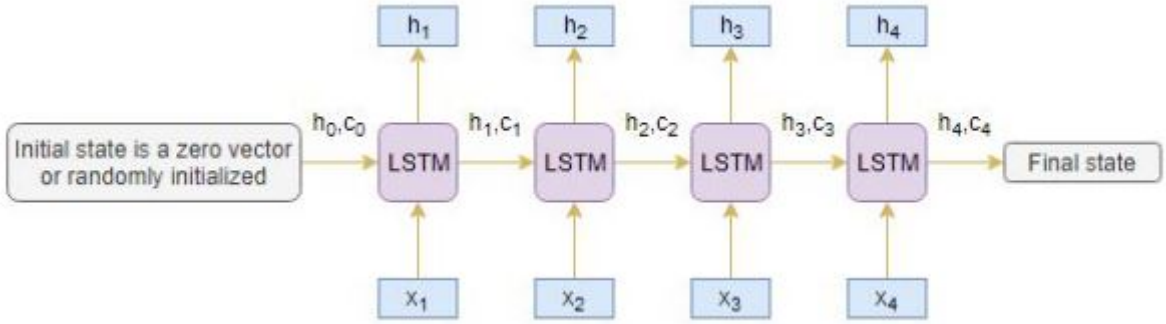


Fig 3.5: Encoder

The hidden state (h_i) and cell state (c_i) of the last time step are used to initialize the decoder. Remember, this is because the encoder and decoder are two different sets of the LSTM architecture.

ii. Decoder

The decoder is also an LSTM network which reads the entire target sequence word-by-word and predicts the same sequence offset by one time step. The decoder is trained to predict the next word in the sequence given the previous word.

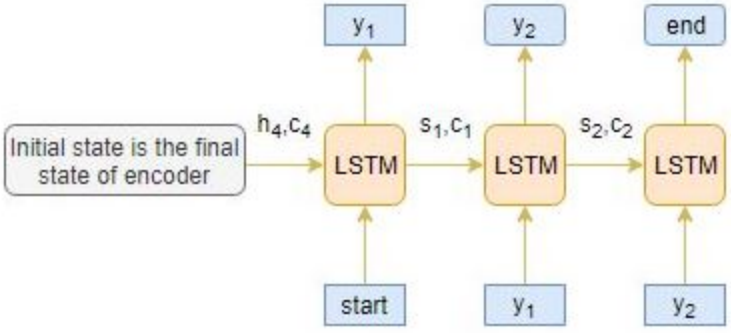


Fig 3.6: Decoder

<start> and <end> are the special tokens which are added to the target sequence before feeding it into the decoder. The target sequence is unknown while decoding the test sequence. So, we start predicting the target sequence by passing the first word into the decoder which would be always the <star**t>** token. And the <end**>** token signals the end of the sentence.

3.3.2.2 Inference Phase

After training, the model is tested on new source sequences for which the target sequence is unknown. So, we need to set up the inference architecture to decode a test sequence. The inference process works as follow

1. Encode the entire input sequence and initialize the decoder with internal states of the encoder
2. Pass <start> token as an input to the decoder
3. Run the decoder for one time step with the internal states
4. The output will be the probability for the next word. The word with the maximum probability will be selected
5. Pass the sampled word as an input to the decoder in the next time step and update the internal states with the current time step
6. Repeat steps 3 – 5 until we generate <end> token or hit the maximum length of the target sequence

3.4 Sentiment Analysis

Sentiment analysis models detect polarity within a text (e.g. a *positive* or *negative* opinion), whether it's a whole document, paragraph, sentence, or clause.

Understanding people's emotions is essential for businesses since customers are able to express their thoughts and feelings more openly than ever before. By automatically analyzing customer feedback, from survey responses to social media conversations, brands are able to listen attentively to their customers, and tailor products and services to meet their needs.

The service and the product review's polarity is the rating the user provides for that review. The Good Reviews are those with rating 5 stars, 4 stars and 3 stars and Bad Reviews are those with rating 2 stars and 1 star. Finally, when a feature sentiment is extracted the sentiment phrase is sent to a polarizer method, this method basically returns 1 if the phrase is a positive sentiment else 0 if the phrase is a negative sentiment.

3.4.1 CountVectorizer

The CountVectorizer provides a simple way to both tokenize a collection of text documents and build a vocabulary of known words, but also to encode new documents using that vocabulary.

It can be done as follows:

1. Create an instance of the *CountVectorizer* class.

2. Call the *fit()* function in order to learn a vocabulary from one or more documents.
3. Call the *transform()* function on one or more documents as needed to encode each as a vector.

An encoded vector is returned with a length of the entire vocabulary and an integer count for the number of times each word appeared in the document. Because these vectors will contain a lot of zeros, we call them sparse.

3.4.2 Predicting Models

Sr.	Algorithm	Accuracy	Precision	Recall
1	Logistic Regression	77.93	79.93	81
2	SVM	75.48	78.03	78.43
3	SVM - Radial Basis	76.07	76.44	82.9
4	Decision Tree	70.60	74.55	72.3
5	Random Forest	63.96	61.41	96.3

Table 3.1: Model Comparison

CHAPTER 4

Results

The main aim of our system is to ensure fair results of sentiments, also we don't want users to spend a lot of time reading through long textual descriptions in the reviews. A user when looking at the reviews of a particular may come across a lengthy feedback which would be tedious for them to go through. Therefore, instead of going through the entire length of the review, the user can easily copy the review and paste it in the following dialogue box.

It provides the user with an option to summarize the entered review as well as analyse it to know more about the previous customer's polarity in regards to the product.

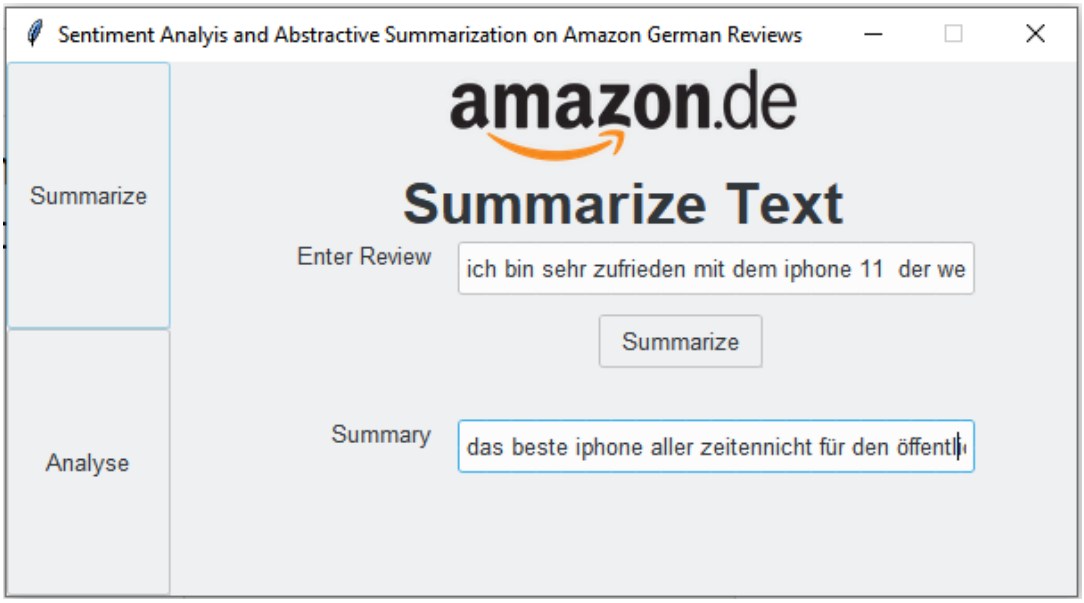


Fig 4.1: GUI(a)

Fig 4.2: GUI(b)

CHAPTER 5

Conclusion and Future Work

Sentiment analysis is a field of study that analyzes people's sentiments, attitudes, or emotions towards certain entities. This project tackles a fundamental problem of sentiment analysis, sentiment polarity categorization. Online product reviews from Amazon.de are selected as data used for this project. Summarization & Sentiment Analysis of German Amazon Reviews successfully creates a summary of long reviews of products posted by buyers on Amazon and performs sentiment analysis to provide future buyers an insight into how the customers have liked the product. It detects the polarity of the review and classifies it as either positive or negative and summarizes the review in an abstractive fashion allowing the user to make appropriate decisions about the product.

REFERENCES

[1] Y. Xu, X. Wu, and Q. Wang. Sentiment analysis of yelps ratings based on text reviews, 2015.

[2] M. S. Elli and Y.-F. Wang. Amazon reviews, business analytics with sentiment analysis. Year : 2013 10th IEEE International Conference on Control and Automation (ICCA) Hangzhou China, June 12-14-2013 .

[3] C. Rain. Sentiment analysis in amazon reviews using probabilistic machine learning. Swarthmore College, 2013.

[4] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. Journal of Machine Learning Research, 12(Aug):2493–2537, 2011.

[5] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 conference on empirical methods in natural language processing, pages 1631–1642, 2013.

[6] Yadav M.P., Feeroz M. and Yadav V.K. 2012 Computing Communication & Networking Technologies (ICCCNT), 2012 Third International Conference on (IEEE) Mining the customer behavior using web usage mining in e-commerce 1-5 July

[7] Hamouda A. and Rohaim M. 2011 World congress on computer science and information technology (IAENG) Reviews classification using sentiwordnet lexicon January