Visvesvaraya Technological University

BELAGAVI, KARNATAKA

ವಿಶ್ವೇಶ್ವರಯ್ಯ ತಾಂತ್ರಿಕ ವಿಶ್ವವಿದ್ಯಾಲಯ

ಬೆಳಗಾವಿ, ಕರ್ನಾಟಕ

Phase-2 Project Report
on

"Customer Reviews for Product Recommendation using Machine Learning"

Submitted by

ABHISHEK J M	**4JN18IS001**
ASHISH K PASTAY	**4JN18IS014**
CHIRANTHAN P	**4JN18IS021**
SHISHIRA S JOIS	**4JN18IS090**

Under the guidance of

Dr. Samara Mubeen, M.Tech., Ph.D.

Assistant Professor,
Dept. of IS&E,
JNNCE, Shivamogga

Department of Information Science & Engineering
J N N College of Engineering
Shivamogga - 577 204

2021-22

Visvesvaraya Technological University

BELAGAVI, KARNATAKA

ವಿಶ್ವೇಶ್ವರಯ್ಯ ತಾಂತ್ರಿಕ ವಿಶ್ವವಿದ್ಯಾಲಯ

ಬೆಳಗಾವಿ, ಕರ್ನಾಟಕ

Phase-2 Project Report [18CSP83]
on

"Customer Reviews for Product Recommendation using Machine Learning"

Submitted by

ABHISHEK J M	**4JN18IS001**
ASHISH K PASTAY	**4JN18IS014**
CHIRANTHAN P	**4JN18IS021**
SHISHIRA S JOIS	**4JN18IS090**

students of 8th semester B.E. ISE, in partial fulfillment of the requirement for the award of degree of Bachelor of Engineering in Information Science and Engineering of Visvesvaraya Technological University, Belagavi during the year 2021-22.

Under the guidance of

Dr. Samara Mubeen, M.Tech., Ph.D.

Assistant Professor,
Dept. of IS&E,
JNNCE, Shivamogga

Department of Information Science & Engineering
J N N College of Engineering
Shivamogga - 577 204

2021-22

National Education Society ®

J N N COLLEGE OF ENGINEERING

SHIVAMOGGA - 577204.

DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that Project entitled

"Customer Reviews for Product Recommendation using Machine Learning"

Submitted by

1. ABHISHEK J M	4JN18IS001
2. ASHISH K PASTAY	4JN18IS014
3. CHIRANTHAN P	4JN18IS021
4. SHISHIRA S JOIS	4JN18IS090

students of 8th semester B.E. ISE, in partial fulfillment of the requirement for the award of degree of Bachelor of Engineering in Information Science and Engineering of Visvesvaraya Technological University, Belagavi during the year 2021-22.

Signature of Guide

Dr. Samara Mubeen M.Tech., Ph.D.
Assistant Professor,
Dept. of IS&E,
JNNCE, Shivamogga

Signature of HOD

Dr. R Sanjeev Kunte M. Tech, Ph.D.
Professor & Head,
Dept. of IS&E,
JNNCE, Shivamogga

Signature of Principal

Dr. K Nagendra Prasad M.E, Ph.D.
Principal, JNNCE, Shivamogga

1. Examiner _____ **2. Examiner** _____

ABSTRACT

Online Shopping is an upcoming trend than the traditional way of doing shopping. The branded products are obtained at reasonable cost at door step. Henceforth the focus of this project is to classifying customer reviews as either recommendable or non-recommendable using machine learning techniques. This provides an excellent option for customers to filter out “good” and “bad” reviews, the problem with this system is that, there can be lack of authenticity in terms of providing ratings and ordering of reviews. This project is done with two end goals: to automatically classify reviews using the reviews/ratings and to showcase the classified reviews using WordCloud.

The main aim of analysis is to identify the polarity of the data in the Web and classify them. As Sentiment analysis or opinion mining is one of the major tasks of NLP (Natural Language Processing); Sentiment analysis has gained much attention in recent years. This project aims to tackle the problem of sentiment polarity categorization, which is one of the fundamental problems of sentiment analysis. Data used in this study are online product reviews collected from the E-commerce websites namely Amazon or Flipkart. Experiments for review-level categorization are performed with promising outcomes.

ACKNOWLEDGEMENT

On presenting the Project report on "**Customer Reviews for Product Recommendation using Machine Learning"** we feel great to express our humble feeling of thanks to all those who have helped us directly or indirectly in the successful completion of the project work.

We would like to thank our respected guide **Dr. Samara Mubeen**, Assistant Professor, Department of ISE, for her continuous encouragement and guidance.

We would like to thank our project coordinators, **Dr. Samara Mubeen,** Assistant Professor, **Mr. Sharath Kumar S R,** Assistant Professor, **Mr. Chethan G S,** Assistant Professor, **Mr. Girish Mantha**, Assistant Professor, Department of ISE, for all their support and encouragement.

We would like to thank **Dr. R Sanjeev Kunte**, Professor and Head of Dept. of ISE, JNNCE, Shimoga and **Dr. K Nagendra Prasad**, Principal JNNCE, Shimoga for all their support and encouragement.

We are grateful to Department of Information Science and Engineering and our institution J N N College of Engineering and for imparting us the knowledge with which we can do our best.

Finally, we would like to thank the whole teaching and non-teaching staff of Information Science and Engineering Department.

Thanking You

ABHISHEK J M - 4JN18IS001
ASHISH K PASTAY - 4JN18IS014
CHIRANTHAN P - 4JN18IS021
SHISHIRA S JOIS - 4JN18IS090

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1:

INTRODUCTION

1.1 Preamble

Today, digital reviews play a pivotal role in enhancing global communications among consumers and influencing consumer buying patterns. E-commerce giants like Amazon, Flipkart, etc. provide a platform to consumers to share their experience and provide real insights about the performance of the product to future buyers. In order to extract valuable insights from a large set of reviews, classification of reviews into positive and negative sentiment is required. Sentiment Analysis is a computational study to extract subjective information from the text.

An increasingly prevalent trend in the sale of goods is the shift to E-commerce, or online shopping. Numerous, if not most, traditional "brick and mortar" stores have online shops where consumers can place orders of many of the same products, they would find at the physical store locations. As this trend continues (often to the disdain of in-store workers), these locations have become simple "showrooms" where customers can see and touch the product, but actually plan to order it online where it may be cheaper, more varied in size or color, or simply more convenient to have shipped rather than brought home. Aside from convenience and competition, the largest benefit to customers is arguably the availability of firsthand reviews and feedback from other shoppers. "What do the reviews say?" and "How many stars did it get?" are questions that online consumer factor in to their purchasing decisions. In addition to the customer benefit, companies making the products being sold also benefit from online availability of such reviews. They can incorporate the feedback of their customers into future product iterations with the end goal of increasing sales. For these reasons, it is of high importance to strive for the best quality and most accurate reviews. One way to judge quality and accurate reviews is by their helpfulness to other readers, which is where this project focuses.

Sentiment is an attitude, thought, or judgment prompted by feeling. Sentiment analysis, which is also known as opinion mining, studies people's sentiments towards certain entities. From a user's perspective, people are able to post their own content through various social media, such as forums, micro-blogs, or online social networking sites. From a researcher's perspective, many social media sites release their application programming interfaces (APIs), prompting data collection and analysis by researchers and developers. However, those types of online data have several flaws that potentially hinder the process of

sentiment analysis. The first flaw is that since people can freely post their own content, the quality of their opinions cannot be guaranteed. The second flaw is that ground truth of such online data is not always available. A ground truth is more like a tag of a certain opinion, indicating whether the opinion is positive, negative, or neutral.

Currently, one of the most popular multi-categories online shop is Flipkart.com. With many products in many different departments, it has become a hugely popular option for online shoppers. This popularity increases the number of customer reviews which in turn adds to the site's utility. Aside from a “star rating” from 1 to 5, customers can also submit textual feedback and product accounts, made available on the one of the product page on Flipkart. Next to each review are three simple user-interface elements: a label, “Was this review helpful to you?”, and two buttons, “Yes” and “No”. It is this mechanism that allows users to vote up or down the helpfulness of a product review. The website then allows customers to sort reviews by their voted helpfulness (the site’s default review ordering) or temporally. While providing an excellent option for customers to filter out “good” and “bad” reviews, the problem with this system is the necessity of participation from review readers and the possibility that reviews that were not voted on or were authored so long ago are not high up in the ordering of reviews. This means that helpful reviews would likely not be seen by customers unless they enumerated through a potentially very large set of other reviews.

To mitigate the issues mentioned above, this project develops a machine learning technique to automatically classify product reviews as recommendable and non-recommendable, using Natural language processing, with two end goals: to automatically classify reviews using the ratings and to showcase the overall review of the product using WordCloud.

1.2 E-commerce

E-Commerce or Electronic Commerce means buying and selling of goods, products, or services over the internet. E-commerce is also known as electronic commerce or internet commerce. These services are provided online over the internet network. Transaction of money, funds, and data is also considered as E-commerce. The standard definition of E-commerce is a commercial transaction which is happened over the internet. Online stores like Amazon, Flipkart, Shopify, Myntra, eBay, Quikr, Olx are examples of E-commerce websites. By 2020, global retail e-commerce can reach up to $27 Trillion. Let us learn in detail about what is the advantages and disadvantages of E-commerce and its types.

E-commerce is a popular term for electronic commerce or even internet commerce. The name is self-explanatory, it is the meeting of buyers and sellers on the internet. This involves the transaction of goods and services, the transfer of funds and the exchange of data.

Types of E-Commerce Models

Electronic commerce can be classified into four main categories. The basis for this simple classification is the parties that are involved in the transactions. So, the four basic electronic commerce models are as follows,

1. Business to Business

This is Business to Business transactions. Here the companies are doing business with each other. The final consumer is not involved. So, the online transactions only involve the manufacturers, wholesalers, retailers etc.

2. Business to Consumer

Business to Consumer. Here the company will sell their goods and/or services directly to the consumer. The consumer can browse their websites and look at products, pictures, read reviews. Then they place their order and the company ships the goods directly to them. Popular examples are Amazon, Flipkart, Jabong etc.

3. Consumer to Consumer

Consumer to consumer, where the consumers are in direct contact with each other. No company is involved. It helps people sell their personal goods and assets directly to an interested party. Usually, goods traded are cars, bikes, electronics etc. OLX, Quikr etc follow this model.

4. Consumer to Business

This is the reverse of B2C, it is a consumer-to-business. So, the consumer provides a good or some service to the company. Say for example an IT freelancer who demos and sells his software to a company. This would be a C2B transaction.

Examples of E-Commerce

Amazon, Flipkart, eBay, Fiverr, Upwork, Olx, Quikr.

Advantages of E-Commerce

1. E-commerce provides the sellers with a global reach. They remove the barrier of place. Now sellers and buyers can meet in the virtual world, without the hindrance of location.

2. Electronic commerce will substantially lower the transaction cost. It eliminates many fixed costs of maintaining brick and mortar shops. This allows the companies to enjoy a much higher margin of profit.
3. It provides quick delivery of goods with very little effort on part of the customer. Customer complaints are also addressed quickly. It also saves time, energy and effort for both the consumers and the company.
4. One other great advantage is the convenience it offers. A customer can shop 24×7. The website is functional at all times, it does not have working hours like a shop.
5. Electronic commerce also allows the customer and the business to be in touch directly, without any intermediaries. This allows for quick communication and transactions. It also gives a valuable personal touch.

Disadvantages of E-Commerce

1. The start-up costs of the e-commerce portal are very high. The setup of the hardware and the software, the training cost of employees, the constant maintenance and upkeep are all quite expensive.
2. Although it may seem like a sure thing, the e-commerce industry has a high risk of failure. Many companies riding the dot-com wave of the 2000s have failed miserably. The high risk of failure remains even today.
3. At times, e-commerce can feel impersonal. So, it lacks the warmth of an interpersonal relationship which is important for many brands and products. This lack of a personal touch can be a disadvantage for many types of services and products like interior designing or the jewellery business.
4. Security is another area of concern. Only recently, we have witnessed many security breaches where the information of the customers was stolen. Credit card theft, identity theft etc. remain big concerns with the customers.
5. Then there are also fulfilment problems. Even after the order is placed there can be problems with shipping, delivery, mix-ups etc. This leaves the customers unhappy and dissatisfied.

M-commerce

Mobile commerce popularly known as m-commerce is actually just a subset of e-commerce. The term itself was coined in 1997 by Kevin Duffy. It is essentially a way of carrying thousands and millions of retail shops in the pocket. Let us study a bit more about mobile commerce.

The use of wireless technology (WAP) to conduct sales of goods, provide services, make payments and other financial transactions, the exchange of information etc. is the basis of mobile commerce.

M-commerce is actually a rapidly growing sector of e-commerce. Nearly 70% of the online transactions that occur in India happen from mobile phones. Globally it is a 700 billion dollar industry.

M-commerce is about exploiting new opportunities made available to us thanks to e-commerce. So, it involves the advent of new technologies, services, business models and marketing strategies. It differentiates itself in many ways from e-commerce. This is because mobile phones have very different characteristics than desktop computers. And it opens so many windows of opportunities for businesses to exploit.

Applications of M-commerce

Other than the straightforward m-commerce transactions of buying and selling of goods and services, they have so many applications. Some applications are listed below:

1. Mobile Banking: Using a mobile website or application to perform all banking functions. It is one step ahead of online banking and has become commonplace these days. For example, in Nigeria, the majority of banking transactions happen on mobile phones.
2. Mobile Ticketing and Booking: Making bookings and receiving tickets on the mobile. The digital ticket or boarding pass is sent directly to your phone after you make the payment from it. Even in India now IRTC and other services provide m-ticketing services.
3. E-bills: This includes mobile vouchers, mobile coupons to be redeemed and even loyalty points or cards system.
4. Auctions: Online auctions having now been developed to be made available via mobile phones as well.
5. Stock Market Reports and even stock market trading over mobile applications.

Advantages of M-commerce

1. It provides a very convenient and easy to use the system to conduct business transactions.
2. Mobile commerce has a very wide reach. A huge part of the world's population has a mobile phone in their pocket. So, the sheer size of the market is tremendous.

3. M-commerce also helps businesses target customers according to their location, service provider, the type of device they use and various other criteria. This can be a good marketing tool.
4. The costs of the company also reduced. This is due to the streamlined processes, now transaction cost, low carrying cost and low order processing cost as well.

Disadvantages of M-commerce

1. The existing technology to set up an m-commerce business is very expensive. It has great start-up costs and many complications arise.
2. In developing countries, the networks and service providers are not reliable. It is not most suitable for data transfer.
3. Then there is the issue of security. There are many concerns about the safety of the customer's private information. And the possibility of a data leak is very daunting.

1.3 Natural Language Processing

Natural language processing (NLP) refers to the branch of computer science and more specifically, the branch of artificial intelligence or AI concerned with giving computers the ability to understand text and spoken words in much the same way human beings can.

NLP combines computational linguistics rule-based modeling of human language with statistical, machine learning, and deep learning models. Together, these technologies enable computers to process human language in the form of text or voice data and to 'understand' its full meaning, complete with the speaker or writer's intent and sentiment.

NLP drives computer programs that translate text from one language to another, respond to spoken commands, and summarize large volumes of text rapidly even in real time. NLP interaction can be done in the form of voice-operated GPS systems, digital assistants, speech-to-text dictation software, customer service chatbots, and other consumer conveniences. But NLP also plays a growing role in enterprise solutions that help streamline business operations, increase employee productivity, and simplify mission-critical business processes.

Basic NLP tasks include tokenization and parsing, lemmatization/stemming, part-of speech tagging, language detection and identification of semantic relationships. In general terms, NLP tasks break down language into shorter, elemental pieces, try to understand relationships between the pieces and explore how the pieces work together to create meaning.

Natural language processing (NLP) is a cutting-edge development for several reasons. Before NLP, businesses were using AI and machine learning for essential insights, but NLP provides the tools to enhance data and analyze both linguistic and statistical data. NLP offers several benefits for companies across different industries.

1.4 Problem Description

In online shopping customer select the product based on the reviews/ratings of the products. If the reviews/ratings itself is not authentic then buying the product online leads to losing the customers from online shopping. This problem is addressed in this project by using machine learning approach and positive and negative reviews are identified.

The application will further provide reviews given by the users and showcase positive and negative reviews using WordCloud, the further aim is to create a recommendation system that recommends products to users based on reviews/ratings.

1.5 Objectives

1. To web scrape product reviews on various websites featuring various products specifically Flipkart.
2. To filter the verified reviews from total number of reviews.
3. To analyze the data and categorize the products based on reviews/ratings for product recommendation
4. To showcase categorized reviews using WordCloud.

1.6 Organization of Report

The further report includes the following contents, Chapter 2 consists of literature survey of different reference papers, Chapter 3 represents system design of the proposed work, Chapter 4 consists results of the proposed work, chapter 5 consists of conclusion and references.

CHAPTER: 2

LITERATURE SURVEY

This chapter presents the related work carried out for Product Recommendation.

A Recommendation System for Online Purchase Using Feature and Product Ranking.

Authors: Karthik.R.V , Sannasi Ganapathy and Arputharaj Kannan.

Description:

Social networks occupy an important place and take a considerable amount of time in people's daily life. It has become so popular that people are sharing a huge amount of data and opinion on social network/review sites, which in turn helps to find interesting insights for organizations / vendors or consumers. In this paper, a new algorithm is proposed called Feature Based Product Ranking and Recommendation Algorithm (FBPRRA) for providing suggestions to the customers who are interested in purchasing good quality products. The proposed algorithm analyzes online products and ranks them according to product reviews. Finally, it recommends the suitable product to the target customers. Experiments have been conducted using online reviews for evaluating the proposed algorithm and found that the proposed recommendation algorithm achieves better prediction accuracy than the existing classifiers such as Naive Bayes, Support Vector Machine, Random Forest, Decision Tree and KNN.

A lot of research is currently happening to extract opinion insights. The collective term for this is known as Sentimental analysis. This paper focuses on the importance of sentimental analysis in social networks and online shopping. Feature selection is used to reduce the dimensionality of the vast data. It detects all the relevant features and discards the irrelevant ones. Feature Selection technique has several benefits: Improves the performance of machine learning, helps in data reduction, reduces computational burden and reduces the storage requirements and space and improves the accuracy of prediction. There are two types of feature selection methods namely filter based approach and wrapper-based approach.

Features expressed in each review are identified and weightage is calculated. In the proposed method a weighted scale is introduced if the review contains both positive and negative, it will not be considered as neutral [1]. This will improve the accuracy of the product recommendations. Using Demographic information and user profile information with Sentiment analysis is a new approach, which is not yet tried as per the knowledge. This proposed algorithm helps to provide the product feature recommendation based on the target

consumer.

Using the pattern extraction method, customer review is checked whether any demographic information target user age group any occasion information is present. This additional information is stored in the database with the product name. Purchaser age group is also stored if user profile information is available, this information is later used for product recommendation. User interested features are also extracted by previous user comments and reviews. If there are no previous comments, then features from similar age groups are selected. Based on the sentiment analysis outcome, ranking is calculated.

Advantages:

1. This feature will be helpful for both customers and the company which produce the product by the reviews obtained.

Disadvantages:

1. Sometimes the reviews and ratings will affect the product ranking which results in poor performance of the system.

Customers' perceptions of online retailing service quality and their satisfaction.

Authors: Jun, M., Yang, Z. and Kim, D.

Description:

Online service quality is one of the key determinants of the success of online retailers.

This exploratory study revealed some important findings about online service quality. First, the study identified six key online retailing service quality dimensions as perceived by online customers: reliable/prompt responses, access, ease of use, attentiveness, security, and credibility [2]. Second, of the six, three dimensions, notably reliable/prompt responses, attentiveness, and ease of use, had significant impacts on both customers' perceived overall service quality and their satisfaction. Third, the access dimension had a significant effect on overall service quality, but not on satisfaction. Finally, this study discovered a significantly positive relationship between overall service quality and satisfaction. Important managerial implications and recommendations are also presented.

Online retailers heavily involve non-human interactions between customers and online retailers' information systems. There are mainly two types of interactions over the Internet:(1) The interactions between customers and online retailers' employees via either Internet-based communication tools, such as e-mail, chat room, and message board, or traditional communication channels. (2) The interactions between customers and online retailers' Web sites, through which customers can search and retrieve necessary information, and place their orders.

Another important aspect of online systems is to enable customers to function more independently and conduct many transactions on their own. As end-users, consumers often seek desired products/services information through Web sites. Thus, in this online market, customers are essentially "self-served" much of the time. Recently, several studies on E-commerce have noted that some features of Web sites are critical to their business success relationship between online retailers' service quality and their customer satisfaction. Service quality improvement initiatives should begin with defining the customers' needs and preferences, and their related quality dimensions. If online retailers understand what dimensions customers use to judge quality, they can take appropriate actions to monitor and enhance performance on those dimensions and remedy service failures. Service quality improvement initiatives should begin with defining the customers' needs and preferences, and their related quality dimensions. If online retailers understand what dimensions customers use to judge quality, they can take appropriate actions to monitor and enhance performance on those dimensions and remedy service failures.

There is a strong and positive relationship between online retailers' service quality and their customer satisfaction.

Advantages:

1. Saves time and effort. Convenience of shopping at home.
2. Wide variety / range of products are available / good discounts / lower prices.
3. Get detailed information about the product.
4. Various models / brands can be compared.

Disadvantages:

1. It is generally not easy for online retailers to gain and sustain competitive advantages based solely on a cost leadership strategy in rival-driven online retailing

Sentiment Analysis on Customer Feedback Data: Amazon Product Reviews

Authors: Pankaj, Prashant Pandey, Muskan and Nitasha Soni.

Description:

Sentiment analysis is one of the fastest spreading research areas in computer science, making it challenging to keep track of all the activities in the area. Sentiment analysis presents customer feedback reviews on products, which utilizes opinion mining, text mining and sentiments, which have affected the surrounding world by changing their opinion on a specific product. Data used in this study are online product reviews collected from Amazon.com. A comparative sentiment analysis of retrieved reviews is performed. This research paper provides sentimental analysis of various smartphone opinions on smartphones dividing them Positive, Negative and Neutral Behavior [3]. Opinions are statements that reflect people's perception or sentiment. Sentiment analysis is a series of methods, techniques, and tools about detecting and extracting subjective information, such as opinion and attitudes, from language, helping in finding the mood of the customers about a purchasing of a particular product or topic. It involves building a system to collect and examine opinions about the product made in many online purchasing sites. Sentiment analysis is a sub field of web content mining. Sentiment analysis has been about opinion contradiction, i.e., whether someone has a positive, neutral, or negative opinion towards something. Data used in this paper is a combination of product reviews collected from Amazon.com. The whole process includes summarization in three steps: (1) Product feature based, which is given by customer; (2) In each review, identify expected features in each opinion sentence and (3) Find out whether the feature/opinion is positive, negative or neutral and finally a summary will be created. Sentiment analysis, also known as Opinion mining, is the study of sentiments that determines the judgement of people's opinions, sentiments, evaluations, and emotions in relation to entities such as products, services, organizations, events, topics and their different attributes. Sentiment Analysis or opinion mining is a case study which analyses people's sentiments, attitudes, entropy or emotions towards certain entities. This paper tackles a fundamental problem of sentiment analysis, sentiment polarity categorization. The data for this research is collected from online product reviews from Amazon.com.

These steps consist of pre-processing, pre-filtering, biasing, data accuracy etc. features which require the knowledge of machine learning. A lot of work in opinion mining

and sentiments of customer reviews has been conducted to mine opinions in form of document, sentence and feature level sentiment analysis. For future preferences, Opinion Mining can be carried out on a set of discovered feature expressions extracted from reviews and become a most interesting research area. There are more innovative and effective techniques that have to be invented which should overcome the current challenges faced by Opinion Mining and sentiment analysis.

Sentiment analysis deals with the classification of texts based on the sentiments they contain. This article focuses on a typical sentiment analysis model consisting of three core steps, namely data preparation, review analysis and sentiment classification, and describes representative techniques involved in those steps. Sentiment analysis is an emerging research area in text mining and computational linguistics, and has attracted considerable research attention in the past few years. Future research shall explore sophisticated methods for opinion and product feature extraction, as well as new classification models that can address the ordered labels property in rating inference. Applications that utilize results from sentiment analysis are also expected to emerge in the near future.

Advantages:

1. Sentiment analysis is a useful tool for any organization or group for which public sentiment or attitude towards them is important for their success - whichever way that success is defined.
2. The results from sentiment analysis help business understand the conversations and discussions taking place about them, and help them react and take action accordingly.

Disadvantages:

1. Computer programs have problems recognizing things like sarcasm and irony, negations, jokes, and exaggerations - the sorts of things a person would have little trouble identifying. And failing to recognize these can skew the results.

Sentiment Analysis Based on Multiple Reviews by using Machine learning approaches

Authors: Stephina Rodney D'souza and Kavita Sonawane.

Description:

Sentiment Analysis can be defined as the process of analyzing online pieces of writing to determine the emotional tone they carry. With the vast growth of social media content on the Internet in the past few years, people now express their opinion on almost anything in

discussion. With respect to this, Bag–of–Words (BoW) is the most popular way to model text in statistical machine learning (ML) approaches [4]. However, the performance of Bow sometimes remains unlimited due to some fundamental deficiencies in handling the polarity shift problem and other few challenges like quality of the opinions, hidden state representations, polarity categorization etc. To come across these challenges the focus will be on Dual Sentiment Analysis which processes the Sentiment with all the perspectives (positive, negative or neutral). This may lead towards the accurate prediction for final decision making based on the reviews given by the customers. The proposed work is being experimented on the Amazon Product reviews specifically the Mobile device reviews. This work aims at overcoming the limitation of existing systems and improving the accuracy.

The research study of implementing the Dual Sentiment Analysis (DSA) seems to be very attractive and interesting in the field of machine learning, as it helps to overcome the challenges such as the Bow model, Polarity shift categorization problems etc. Experimentation of the proposed work with Amazon product data, specifically the Mobile device reviews with the intention of overcoming the limitations of existing systems and improving the accuracy is researched in DSA. It has proved the same in the results obtained. Based on the result and analysis we can conclude that DSA contributed positively in the final decision-making process that would help the customer to gain the correct insights about the reviews; in turn to choose or recommend the correct product.

Advantages:

1. By using sentiment analysis, customers feel about different areas of the business without having to read thousands of customer comments at once.
2. If there are thousands of feedbacks per month, it is not impossible for one person to read all of these responses. By using sentiment analysis and automating this process, they can be can easily drill down into different customer segments of your business and get a better understanding of sentiment in these segments.

Disadvantages:

1. While sentiment analysis is useful, it is not a complete replacement for reading survey responses. Often, there are useful nuances in the comments themselves.

The Impact of Review Environment on Review Credibility

Authors: Jo Mackiewicz, Dave Yeats, Thomas Thornton.

Description:

Increasingly, professional and technical communicators analyze, synthesize, and respond to

user-generated content, including online consumer reviews of products, as the influence of user-generated content on consumers' purchasing decisions grows. But product reviews vary in the degree to which people perceive them to be credible [5].

Testing the effect of a consumer review's environment (brand or retailer site) and the effect of review valence (positive or negative) on the perceived credibility of that review, as well the degree of correlation among credibility, trustworthiness, and expertise. Through an online survey, respondents are exposed to the same review text with different star ratings (4-star and 2-star) in two types of sites: brand and retailer. Participants are asked to evaluate the review's credibility, trustworthiness, and expertise. In half of the exposures, participants evaluated a review in the site of a high-credibility company (Apple or Amazon), and in the other half of exposures, participants evaluated a review in the site of a mid-level-credibility company (Dell or Walmart).

Credibility strongly correlated with both trustworthiness and expertise. Participants rated 4-star reviews as more credible than 2-star reviews on high-credibility sites, but star ratings had no impact on mid-level credibility sites. There was no difference between ratings of reviews displayed on brand and retailer sites for mid-level-credibility companies but a small difference between reviews displayed on brand and retailer sites for high-credibility companies.

Advantages:

1. Once a website scraping service starts collecting data, the data is not only obtained from a single page but from the whole domain.
2. Simple errors in data extraction can lead to major issues. Hence it is needed to ensure that the data is correct. Data scraping is not only a fast process, but it’s accurate too. This reputation helps while collecting important data such as sales price, financial data to name a few.

Disadvantages:

1. This study reflects a limitation of studying the effects of a single review's environment as opposed to accounting for the possible influence of other reviews for the same product.
2. Further research certainly could test the degree to which a review norm with other reviews of the same product. Such a study would mimic a real-life use situation in that review users typically read more than one review when trying to make a purchasing decision.

Keyword extraction from Tweets using NLP tools for collecting relevant news

Authors: Thiruni D. Jayasiriwardene, Gamage Upeksha Ganegoda

Description:

Keywords play a major role in representing the gist of a document. Therefore, a lot of Natural Language processing tools have been implemented to identify keywords in both structured and unstructured texts. Text that appears in social media platforms such as twitter is mostly unstructured because of the character limitation. Consequently, a lot of short terms and symbols such as emoticons and URLs are included in tweets. Keyword extraction from grammatically ambiguous text is not easy compared to structured text since it is hard to rely on the linguistic features in unstructured texts. But when it comes to news on twitter, it may contain somewhat structured text than informal text does but it depends on the tweeter, the person who posts the tweet. In this paper, a methodology is proposed to extract keywords from a given tweet to retrieve relevant news that has been posted on twitter, for fake news detection. The intention of extracting keywords is to find more related news efficiently and effectively. For this approach, a corpus that contains tweet texts from different domains is built in order to make this approach more generic instead of making it a domain- specific approach. In fact, the Stanford Core NLP tool kit, Wordnet linguistic database and statistical method are used for extracting keywords from a tweet. For the system evaluation, the Turing test which has human intervention is used. The system was able to acquire an accuracy of 67.6% according to the evaluation conducted.

This paper proposes a methodology to extract keywords from a given tweet text for the purpose of retrieving relevant news that has been posted on twitter to collect data for fake news detection. The proposed method uses Stanford core NLP, POS tagging, NER as well as TF-IDF statistical method for keyword extraction. In Addition, Wordnet lexical database has been used to find synonyms and Ginsim can be used along with word2vector to find synonyms for the words and the amount of similarity of words. Then the bi-gram technique is used to generate key phrases to increase the accuracy and efficiency for retrieving relevant news. Extracted keywords are used to gather the most relevant news tweets for the claimed tweet. In fact, the set of tweets retrieved were filtered and duplicates were removed to get a clean set of tweets to help detect fake news. The Evaluation can be done using the Turing test and more attention should be paid to the standards of the

participants of the test.

Advantages:

1. NLP system offers exact answers to the questions, no unnecessary or unwanted some information.
2. The accuracy of the answer increases with the amount of relevant information provided in the questions.
3. Structuring a high unstructured data source.

Disadvantages:

1. NLP system doesn't have a user interface that lacks features that allow users to further interact with the system.
2. If it is necessary to develop a model with a new one without using a pre-trained model, it can take a week to achieve a good performance depending o the amount of data.

Aspect Based Summarization of Reviews Using Naïve Bayesian Classifier and Fuzzy Logic

Authors: Reshma V and Ansamma John.

Description:

E-commerce is getting popular; more and more products are sold online every day and product reviews are growing rapidly. The larger number of reviews makes it difficult for customers and manufacturers. For popular products there may have thousands of reviews. Customers and manufacturers may not be able to understand overall opinion about aspects of products without going through all reviews. Also, most of the existing methods make use of approximate summarization of product reviews [7]. Here a more precise and realistic value of opinion is retrieved through naive Bayesian classifier and fuzzy method. In this, in addition to the identification of opinions, linguistic hedges are identified and apply fuzzy rules to magnify the effect of opinion.

In this era, most of the people are depending on internet whether it is for business, education, entertainment or any other purpose. Thus, internet has become an integral part of our life. Now a day, we have witnessed tremendous development in the field of e-commerce. Millions of products are now available to users from merchants of different parts of world. Due to wide variety of products, attractive offers with fast and secure transaction, more and more people are attracted towards this online shopping. Also, people are free to express their opinions and feedbacks on various products through different social media's.

In machine learning, Naive Bayesian classifiers are statistical classifiers which predict the membership property in terms of probability and works with the naïve independence assumption. The main reason for choosing Bayesian classifier is that it uses a very simple and intuitive method. The classification is done on large amount of diverse data. So, time taken for training and testing is important. The time needed to train this model is less compared to other machine learning algorithms such as support vector machine classifier. Bayesian classifiers have also exhibited good accuracy compared to other machine learning techniques.

A simple method of association rule mining and probabilistic approach gave important aspects with very high accuracy. The sentimental analysis using SentiWordNet, naïve Bayesian classifier and fuzzy logic gave more precise and realistic scores for opinions. The experimental result shows that, with sufficient training the system is able to give reliable results.

Advantages:

1. This algorithm works very fast and can easily predict the class of a test dataset.
2. Naive Bayes classifier performs better than other models with less training data if the assumption of independence of features holds.
3. If there are categorical input variables, the Naive Bayes algorithm performs exceptionally well in comparison to numerical variables.

Disadvantages:

1. It assumes that all the features are independent. While it might sound great in theory, in real life and independent features are hardly found in the set.

Classification-based Adaptive Web Scraper

Authors: Ujwal B V S, Bharat Gaind, Abhishek Kundu, Anusha Holla, Mukund Rungta.

Description:

Web scraping is an important problem in computer science. The problem with the commonly-used position or structure-based web scraping tools is that they need to be manually reconfigured as soon as the structure of the web page changes. In this paper, problem of information extraction for web pages consisting of repetitive blocks are performed. Extraction of these blocks and their constituent attributes, using a novel classification-based approach. This approach gives high accuracy when used to extract product-offers from an offer-aggregator website. It is also highly adaptive to the changing structure of a website [8].

Web scraping is an ever-growing, important problem in computer science, with a

plethora of applications in both business and research domains. In many such applications, it is very important to extract lists and its constituent attributes from a web page. Building a product-offer-based recommendation system, which scrapes its content from a variety of websites (containing lists of offers) and recommends the best deals to users, is one such example. Secondly, web pages are susceptible to change, simply because the look-and-feel of the website needs to evolve with time. Due to this, the commonly-used scraping tools need to be manually reconfigured every time, since they are position or structure-based (and use CSS Selectors or xpaths).

This approach is tested by applying it to the specific problem of extracting product-offers from various web pages of an offer-aggregator website and achieved remarkable results. Then, the older versions of these web pages are tested, having completely different structures and still got a high accuracy, which proves our claims experimentally as well. Some other useful applications of our approach could be extracting product-information from E-commerce websites; flight-information from travel company websites; or real-estate listings.

Advantages:

1. Cost-Effective: Web scraping services provide an essential service at a competitive cost.
2. Low Maintenance and Speed: Web Scraping does have a very low maintenance cost associated with it over a while.
3. Data Accuracy: Simple errors in data extraction can lead to major issues.

Disadvantages:

1. Data Analysis of Data Retrieved through Scraping the Web. To analyze the retrieved data, it needs to be treated first.
2. Difficult to Analyze. For those who are not much tech-savvy and aren't an expert, web scrapers can be confusing.

Identifying Deceptive Reviews Using Networking Parameters

Authors: Tanya Gera, Deepak Thakur and Jaiteg Singh.

Description:

Nowadays, client likes to take suggestions before spending on a new product. For this client goes to online item review webpage for perusing other's encounters and saying for that item. A real issue which was disregarded so far is the investigation of review spammers. However, numerous scientists gave their productive commitment in this field of exploration from 2007. The situation now asks for, conspicuous verification and ID of fake reviews and fake reviewers; as this has transformed into a colossal social issue. Those studies have the limit

perceive certain sorts of spammers, in any case, in fact, there are distinctive sorts of spammers who can control their practices to act much the same as certified users [9]. This has transformed into a gigantic social issue. From various years, email spam and web spam were the two essential highlighted social issues.

In the meantime, nowadays, on account of reputation of customers' energy to web shopping and their dependence on the online reviews, it transformed into a true center for review spammers to misdirect customers by making sham overviews for target things. To the best of our insight, very little study is accounted for in regards to this issue reliability of online reviews. In the past few years, variety of techniques has been recommended by researchers to accord with this trouble. This paper intends to identify suspicious review, review spammers and their group using rule-based classification methods along with networking parameters.

Web has been ceaselessly giving imperative wellspring of conclusions on things, organizations, events, individuals etc. Various researchers have helped in the field of sentiments extraction. In any case, a huge part of the work has been fixated using data mining techniques. A basic and veritable issue that has been neglected so far is review spam or steadiness of those online reviews. These activities are performed by some online characters that trick others and the term used for them is sock puppeting. It is the most serious threat to the overall population. Befuddle comments benefits some individuals. This, shockingly, has transformed into an enormous wellspring of compensation for some thought spammers at the cost of deliberately offering out the customers.

All around, there are three sorts of spam- Web spam, email spam and review spam. The purpose of Web spam is to appeal the people to visit some target pages and thusly raising the rank of those. A substitute kind of spam is email spam, which is furthermore not the same as review spam. Email spam (furthermore called rubbish messages) incorporates getting the uninvited business plugs.

In the past few years, the major highlighted field of research was sentiment analysis. These studies assume all the reviews to be genuine. However, due to increase in demand of online shopping, spam has become a big social issue. It is important to identify and detect those review spam. This works intends to produce experimental results of study of impact of networking parameters over rule-based classification method to identify suspicious reviews on shopping websites.

Advantages:

1. It makes everything noticeable.
2. Transparency.

Disadvantages:

1. Purchasing the network cabling and file servers can be expensive.
2. Managing a large network can be complicated.

Summarizing Customer Review Based on Product Feature and Opinion

Authors: Jawad khan, Byeong Soo Jeong

Description:

Opinion or sentiment analysis has risen to extract useful information from a lot of unstructured text data, in the form of customer reviews on different products and their features or online SNS data respectively. Customer reviews are not only helpful for potential customers, but also are helpful for the manufacturers of the products to raise their products and services. The reviews conciseness takes the attention of the customers rather than long reviews. Opinion Mining is playing a major role to summarize customer reviews and make it easy for online customers to determine whether to purchase the products or not. In this paper, supervised lazy learning model utilizing syntactic rules for the product features and opinion words extraction in subjective review sentences are proposed. In the lazy learning algorithm, i.e., K-NN with k=3 is used for the review sentences' classification into two classes (subjective, objective).

Experiment shows that proposed method can improve the performance of existing work in terms of average precision, recall and f-score for the extraction of opinion sentences and product features [10].

Opinion Mining is the computational study which extracts people's opinions, sentiments, perceptions, and emotions about entities, events and their properties. Nowadays, online shopping has become popular as mostly people prefer to buy products online. Before purchasing, customers read other people's reviews on products and their features, and want to collect information from it. These reviews are long and unstructured, so it is inconvenient for a potential customer to read all the reviews and make a decision for product purchasing. Similarly, it is also inconvenient for a manufacturer to know customer opinions about their products in order to develop marketing strategy and product placement in the market basket.

Using linguistic patterns, product features and opinion words are successfully extracted from subjective review sentences. SentiWordNet is used for the orientation of opinion sentences determination based on the high score value of opinion words. The proposed method improved results of average precision, recall and f-score as compared to the existing work.

Review summary is created based on product features, which provide fast insights to customers and manufacturers, easy review of products and time saving. The method has to be further improved by including implicit features and determining the strength of opinion.

Advantages:

1. Since Natural language processing and data mining are used, data obtained will be more accurate.

Disadvantages:

1. Data is classified as only positive and negative. It will be difficult to analyze the product which is at the average rating.

Sentiment Analysis Using Latent Dirichlet Allocation and Topic Polarity WordCloud Visualization

Authors: Mohammad F. A. Bashri, Retno Kusumaningrum.

Description:

Sentiment analysis is a field of study that analyzes sentiment. One method for doing sentiment analysis is Latent Dirichlet Allocation (LDA) that extracts the topic of documents where the topic is represented as the appearance of the words with different topic probability. Therefore, we need data representation in visual form that is easier to understand than text and tables. One form of data visualization is WordCloud that provides a visual representation of words frequency. This research will perform sentiment analysis from the students' comments toward a university, in this case the Universitas Diponegoro, using LDA and topic polarity WordCloud visualization. The purpose of this study is to generate the topic polarity WordCloud of the students' comments by using the best combination of parameters. The best combination is the parameter with the value of alpha 0.1, value of beta 0.1, number of topics 9, threshold 10-7, and perplexity values 8.07. Such parameter combination produces 3 topics as positive sentiment and 6 topics as negative sentiment. In addition, we also compare the proposed method to several algorithms such as Naïve Bayes and Logistic Regression. The final result shows that the proposed method outperforms the Naïve Bayes and Logistic Regression in terms of F-Measure by 61%, 54%, and 56%, respectively [11].

Currently, opinion has a major role to every human action. When someone wants to make a decision, he needs to know what the other people's opinions are about what he will do. For example, when a person wants to buy a product from an online store, he'll look at the reviews from other users to decide whether to buy the product or not. For institutions that engaged in the field of goods and services, they need consumer's opinion regarding the goods or services they produce.

Sentiment analysis, also called opinion mining, is the field of study that analyses people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as

products, services, organizations, individuals, issues, events, topics, and their attributes. Sentiment analysis has several benefits such as product monitoring that is to monitor the level of satisfaction with the education system applied. Several methods have been implemented to analyse sentiment on Indonesian documents, such as Naive Bayes classifier, SVM, and Maximum Entropy.

Advantages:

1. It reveals the essential. Brand names pop and key words float to the surface.
2. They delight and provide emotional connection. Both the creation of a word-cloud and the observation of one help to provide an overall sense of the text. The same visceral response doesn't happen when staring at a page of text.
3. They're fast. Poring over text to develop themes from research takes time.

Disadvantages:

1. Size isn't everything. Although the Word Cloud is designed to make words stand out according to their size based on their frequency of occurrence, other factors can affect the visual 'decoding' of the data from the observer's perspective. For example, the length of the word and the white space around the glyphs (letters) can make it look more or less important relative to others in the cloud. This can mislead your interpretation.

Summary:

This chapter is about the different papers referred, in order to carry out the project work.

CHAPTER: 3

SYSTEM DESIGN AND IMPLEMENTATION

This chapter presents the details of the proposed methodology for developing and implementation of the recommendation framework for analyzing the quality of product based on online ratings/reviews using machine learning.

3.1 Proposed System Design

Figure 3.1 Shows the proposed system design for the recommendation system.

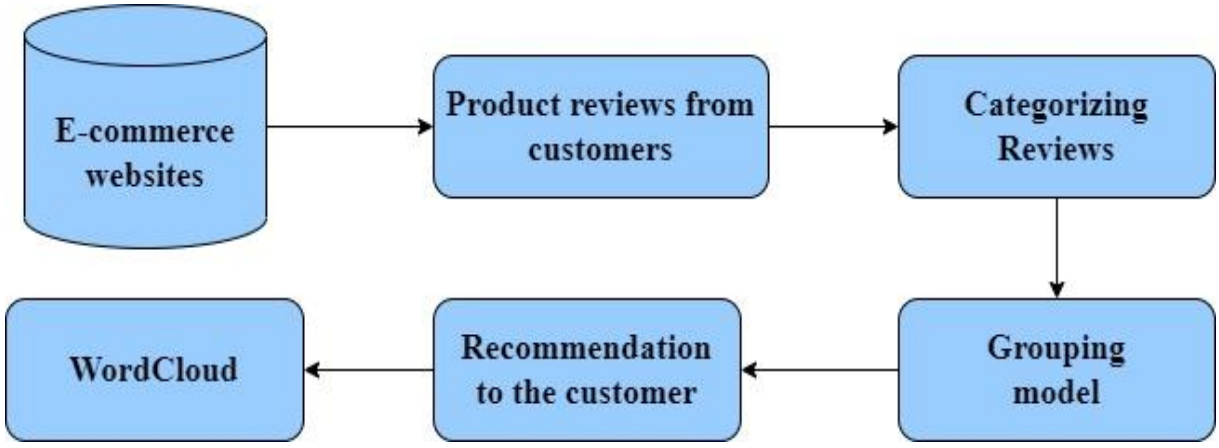

Figure 3.1: Proposed system design

The following are the steps in proposed system design

E-commerce Websites

An E-commerce website, by definition, is a website that allows you to buy and sell tangible goods, digital products or services online.

In this project we are mainly focusing on the major E-commerce website that is Flipkart.com. The particular product is fetched in these E-commerce websites.

Product Reviews from Customers

When the Product is found the next step is to get the ratings and reviews. Here the reviews and ratings that are given by the users on particular products are collected by web scraping using Beautiful Soup.

Categorizing the reviews

Here the reviews are categorized based on ratings as positive if the ratings are one, two and three, and negative if the ratings are four, five.

Grouping Model

Grouping model is the step where all the categorized reviews and ratings on the particular product are grouped together and sentiment analysis on the reviews is done.

Recommendation to the customer

If the positive reviews are more than sixty percent (threshold) of the number of reviews analyzed then the product is recommended or else the product is not recommended.

WordCloud

Finally, the reviews are visually represented by a concept called WordCloud using NLP. A word cloud is a collection, or cluster, of words depicted in different sizes. The bigger and bolder the word appears, the more often it's mentioned within a given text and the more important it is.

3.2 TOOLS USED

This section presents the tools used during implementation.

3.2.1 Web Scraping

Web scraping is an automatic method to obtain large amounts of data from websites. Most of this data is unstructured data in an HTML format which is then converted into structured data in a spreadsheet or a database so that it can be used in various applications. There are many different ways to perform web scraping to obtain data from websites. These include using online services, particular API's or even creating your code for web scraping from scratch. Many large websites, like Google, Twitter, Facebook, Stack Overflow, etc. have API's that allow you to access their data in a structured format. This is the best option, but there are other sites that don't allow users to access large amounts of data in a structured form or they are simply not that technologically advanced. In that situation, it's best to use Web Scraping to scrape the website for data.

Web scraping requires two parts, namely the crawler and the scraper. The crawler is an artificial intelligence algorithm that browses the web to search for the particular data required by following the links across the internet. The scraper, on the other hand, is a specific tool created to extract data from the website. The design of the scraper can vary greatly according to the complexity and scope of the project so that it can quickly and accurately extract the data.

There are mainly two ways to extract data from a website:
1. Use the API of the website (if it exists). For example, Facebook has the Facebook Graph API which allows retrieval of data posted on Facebook.
2. Access the HTML of the webpage and extract useful information/data from it. This technique is called web scraping or web harvesting or web data extraction.
The steps involved in web scraping using the implementation of a Web Scraping framework of Python called Beautiful Soup.

Web Scraping Algorithm:

1.Identify the target website

2.Collect URLs of the pages where you want to extract data from

3.Make a request to these URLs to get the HTML of the page

4.Use locators to find the data in the HTML

5.Save the data in a JSON or CSV file or some other structured format.

3.2.2 Steps involved in Web scraping

1. Send an HTTP request to the URL of the webpage that has to be accessed. The server responds to the request by returning the HTML content of the webpage. For this task, a third-party HTTP library for python-requests is used.
2. Once the HTML content is accessed, the task of parsing the data is left. Since most of the HTML data is nested, the data cannot be extracted simply through string processing. One needs a parser which can create a nested/tree structure of the HTML data. There are many HTML parser libraries available but the most advanced one is html5lib.
3. Now, all that has to be done is to navigate and search the parse tree that is created, i.e., tree traversal. For this task, another third-party python library is used, called Beautiful Soup. It is a Python library for pulling data out of HTML and XML files.

3.2.3 Beautiful Soup

Beautiful Soup is a Python library for pulling data out of HTML and XML files. It works with your favorite parser to provide idiomatic ways of navigating, searching, and modifying the parse tree. It commonly saves programmers hours or days of work.

While the name sounds like something made by a hungry individual, it is, however, a very beautiful tool for web scrappers because of its core features. It can help the programmer to quickly extract the data from a certain web page.

Using Beautiful Soup is not a one-stop solution. To get the most out of it, will few libraries are used. A library is needed to make a request to the website because it can't able to make a request to a particular server. To overcome this issue, it takes the help of the most popular library named Requests or Urlib2. These libraries will help us to make request to the server.

After downloading the HTML, XML data into the local Machine, Beautiful Soup requires an External parser to parse the downloaded data. The most famous parsers are lxml's XML parser, lxml's HTML parser, HTML5lib, HTML. Parser.
Some of the advantages of beautiful soup include:

1. It has good comprehensive documentation which helps us to learn things quickly.
2. It has good community support to figure out the issues that arise while working with the library.

3.2.4 WordCloud

A tag cloud is a visual representation of text data, which is often used to depict keyword metadata on websites, or to visualize free form text. Tags are usually single words, and the importance of each tag is shown with font size or color. When used as website navigation aids, the terms are hyperlinked to items associated with the tag.

Also known as tag clouds or text clouds, these are ideal ways to pull out the most pertinent parts of textual data, from blog posts to databases. They can also help business users compare and contrast two different pieces of text to find the wording similarities between the two.

Perhaps leveraging advanced data visualization techniques is used to turn important analytics into charts, graphs, and infographics. This is an excellent first step, as the brains prefer visual information over any other format.

There are industry tools that allows to code such open-ended data so users can

understand it quantitatively. Yet, these don't come cheap. WordCloud offer a cost-effective, yet powerful, alternative.

Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using a word cloud. Word clouds are widely used for analyzing data from social network websites.

Algorithm

1. Input the string.

2. Tokenize the string into list of words with string tokenizer.

3. Count the words.

4. Sort by the count in descending order.

5. Clean the tokenized words by removing stop words.

6. Assign each word a font size proportional to its count.

7. Generate WordCloud with help of sorted N words.

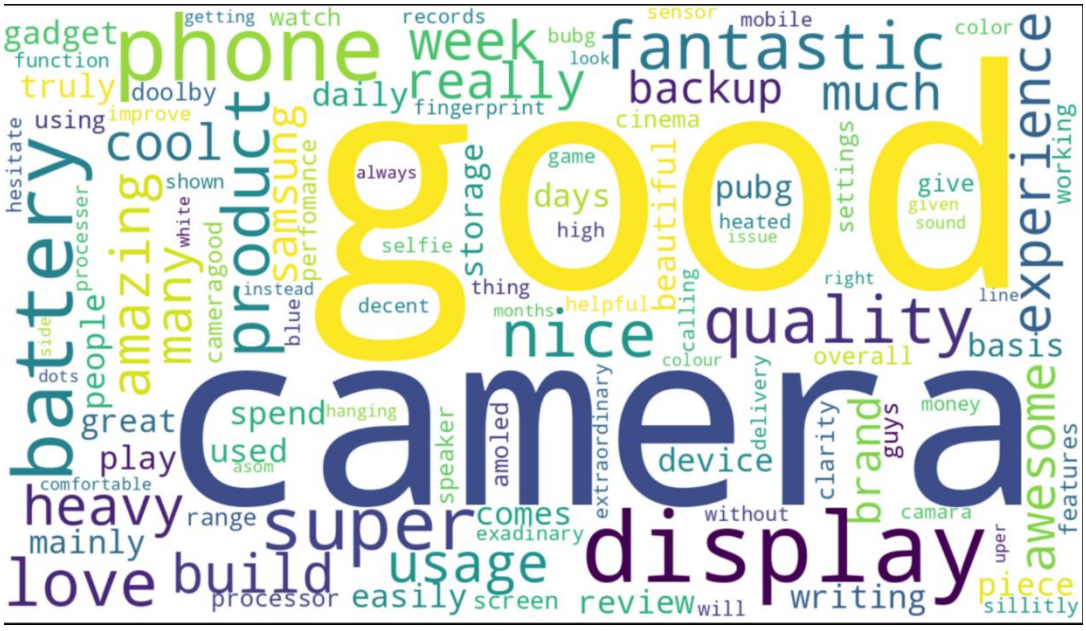

Figure 3.2: WordCloud

3.2.5 Natural Language Toolkit

NLTK, an open-source collection of libraries, programs, and education resources for building NLP programs.

The NLTK includes libraries for many of the NLP tasks listed above, plus libraries for subtasks, such as sentence parsing, word segmentation, stemming and lemmatization (methods of trimming words down to their roots), and tokenization (for breaking phrases, sentences, paragraphs and passages into tokens that help the computer better understand the text). It also includes libraries for implementing capabilities such as semantic reasoning, the ability to reach logical conclusions based on facts extracted from text.

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.

Thanks to a hands-on guide introducing programming fundamentals alongside topics in computational linguistics, plus comprehensive API documentation, NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. Best of all, NLTK is a free, open source, community-driven project.

NLTK has been called “a wonderful tool for teaching, and working in, computational linguistics using Python,” and “an amazing library to play with natural language.”

Natural Language Processing with Python provides a practical introduction to programming for language processing. Written by the creators of NLTK, it guides the reader through the fundamentals of writing Python programs, working with corpora, categorizing text, analyzing linguistic structure, and more. The online version of the book has been updated for Python 3 and NLTK 3.

3.2.6 Flask

Flask is a small and lightweight Python web framework that provides useful tools and features that make creating web applications in Python easier. It is developed by Armin Ronacher who leads an international group of python enthusiasts (POCCO). It gives developers flexibility and is a more accessible framework for new developers can build a web application quickly using only a single Python file.

Python is the most sort after language for web application development and data science as well. It has risen to this height for its ease of use and variety of supportive libraries. There are legacy frameworks like Java's Enterprise edition and ASP. NET's MVC framework is still popular for enterprise-level development. But Python is the favorite for new POC and small-time development where an audience that of an enterprise is not immediately expected. And, of course, the fact that Python and most of its libraries are open sources and free is an exceptionally helpful and useful factor too.

3.2.7 Regular Expression

A regular expression is a sequence of characters that forms a search pattern. When you search for data in a text, you can use this search pattern to describe what you are searching for.

A regular expression can be a single character, or a more complicated pattern.

Regular expressions can be used to perform all types of text search and text replace operations.

3.2.8 OS Module

The OS module in Python provides functions for interacting with the operating system. OS comes under Python's standard utility modules. This module provides a portable way of using operating system-dependent functionality. The os and os.path modules include many functions to interact with the file system.

3.2.9 Joblib Module

Joblib is a set of tools to provide lightweight pipelining in Python. In particular transparent disk-caching of functions and lazy re-evaluation (memorize pattern) easy simple parallel computing. Joblib is optimized to be fast and robust on large data in particular and has specific optimizations for NumPy arrays. It is BSD-licensed.

3.2.10 Requests Module

Requests library is one of the integral parts of Python for making HTTP requests to a specified URL. Whether it be REST APIs or Web Scrapping, requests are must to be learned for proceeding further with these technologies. When one makes a request to a URI, it returns a response. Python requests provides inbuilt functionalities for managing both the request and response.

3.2.11 GitHub

GitHub is a web-based version-control and collaboration platform for software developers. Microsoft, the biggest single contributor to GitHub, initiated an acquisition of GitHub for $7.5 billion in June, 2018. GitHub, which is delivered through a software-as-a-service (SaaS) business model, was started in 2008 and was founded on Git, an open-source code management system created by Linus Torvalds to make software builds faster.

Git is used to store the source code for a project and track the complete history of all changes to that code. It allows developers to collaborate on a project more effectively by providing tools for managing possibly conflicting changes from multiple developers. GitHub allows developers to change, adapt and improve software from its public repositories for free, but it charges for private repositories, offering various paid plans. Each public or private repository contains all of a project's files, as well as each file's revision history. Repositories can have multiple collaborators and can be either public or private.

GitHub facilitates social coding by providing a web interface to the Git code repository and management tools for collaboration. GitHub can be thought of as a serious social networking site for software developers. Members can follow each other, rate each other's work, receive updates for specific projects and communicate publicly or privately.

Three important terms used by developers in GitHub are fork, pull request and merge. A fork, also known as a branch, is simply a repository that has been copied from one member's account to another member's account. Forks and branches allow a developer to make modifications without affecting the original code. If the developer would like to share the modifications, she can send a pull request to the owner of the original repository. If, after reviewing the modifications, the original owner would like to pull the modifications into the repository, she can accept the modifications and merge them with the original repository. Commits are, by default, all retained and interleaved onto the master project, or can be combined into a simpler merge via commit squashing.

GitHub is so intuitive to use and its version-control tools are so useful for collaboration, nonprogrammers have also begun to use GitHub to work on document-based and multimedia projects. GitLab is an open-source alternative to GitHub.

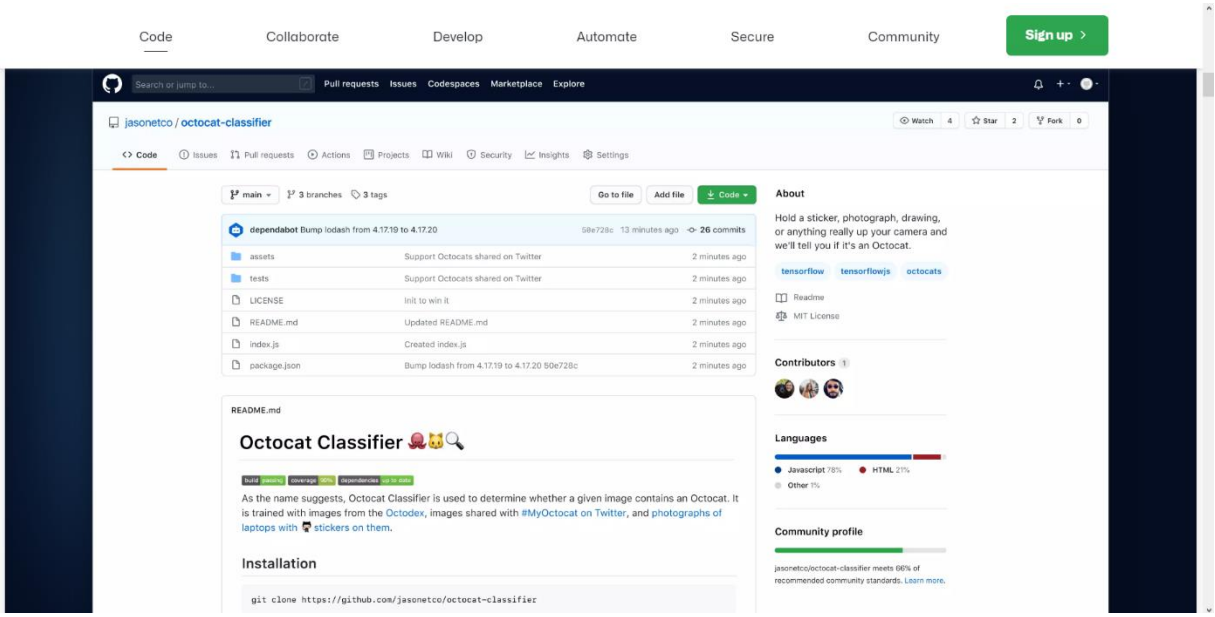

Figure 3.3: GitHub and its Repository

3.3 Use Case Diagram:

Figure 3.3 shows the use case model for Recommendation system.

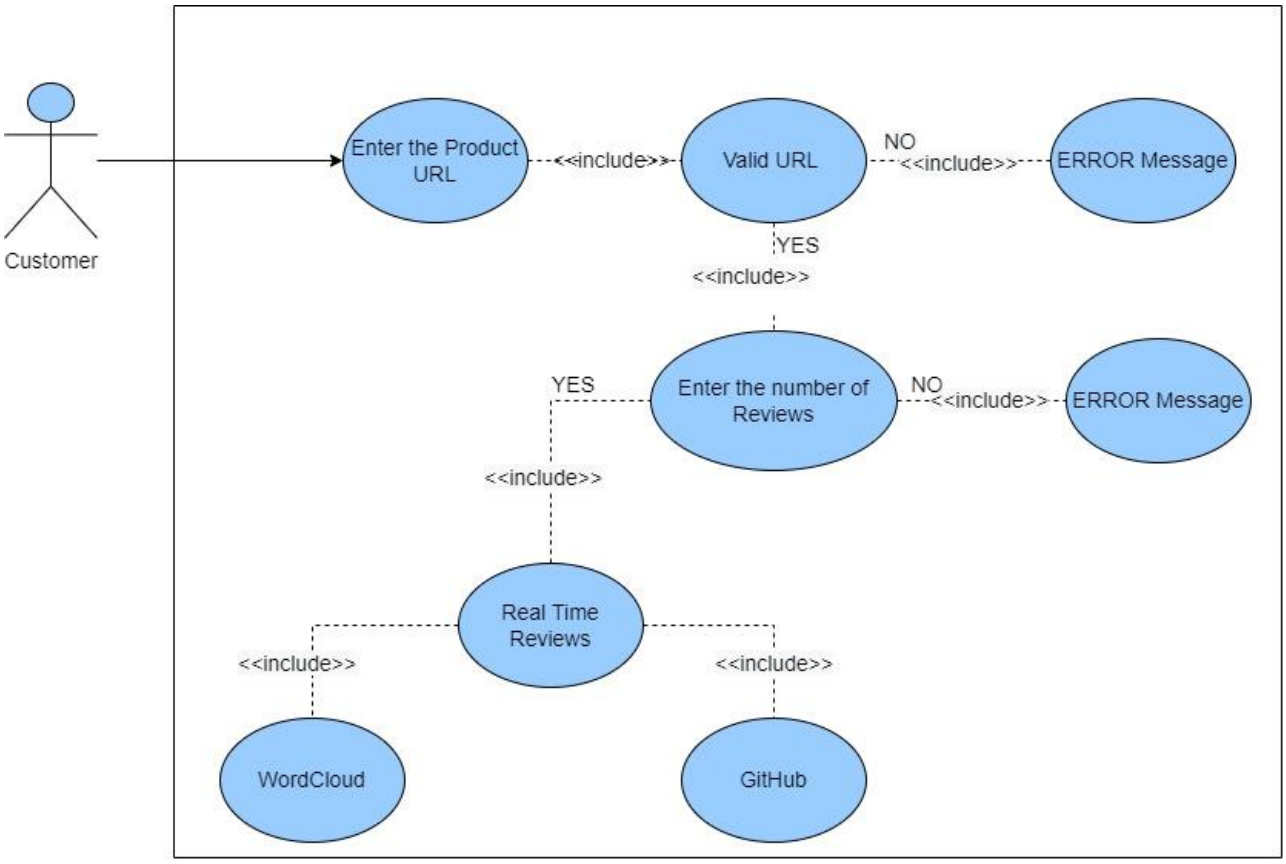

Figure 3.4: Use Case Model of recommendation system

Customer enter the URL of the product to be searched, if the entered URL is Valid, then the number of reviews should be entered, else if it is invalid, it will display error message.

Then if the entered number of reviews is valid, the real time reviews/ratings are categorized and displayed, else if it is invalid, it will display error message.

Further the reviews are visually represented using WordCloud and the source code is available in GitHub.

3.4 Flowchart of the System

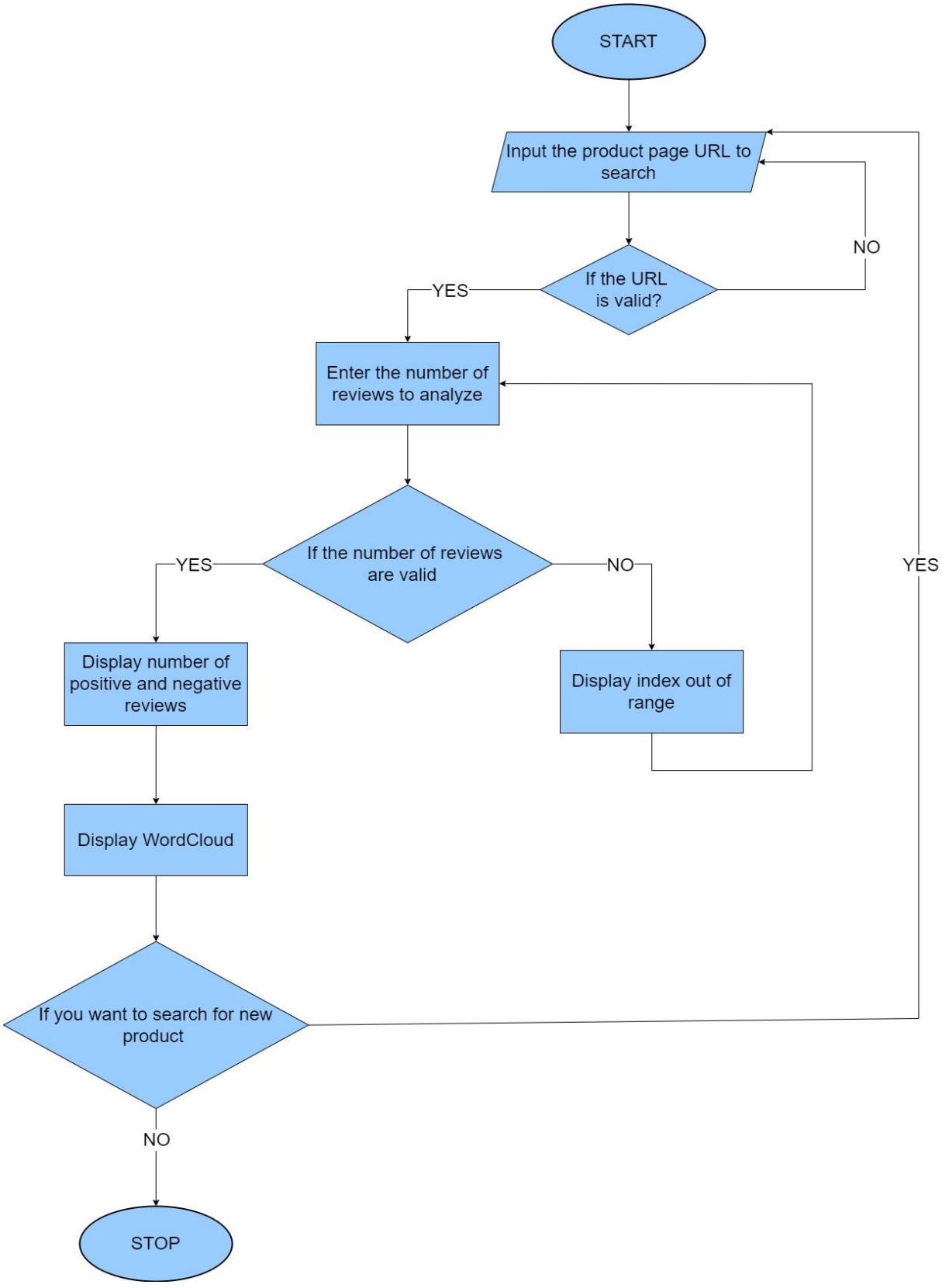

Figure 3.5: Flowchart of the system

3.4.1 Steps involved in Flowchart

Following are the steps involved in flowchart:

Step1: Start

Step2: Input of the Product URL is taken through the search bar.

Step3: If the URL is valid the number of reviews that as to be analyzed is taken from the user, if the URL is not valid the user will be instructed to enter the valid URL.

Step4: If the number of reviews entered is in range, the positive and the negative reviews are displayed with the total number of reviews entered to analyze, if the number is out of range the user has to enter the valid range.

Step5: Display of WordCloud.

Step6: If the user wants to continue and want to search more product, user can visit the home page.

Step7: Stop

3.4.2 System Process

System process contains the information about the input which is given, Internal Process that will take place and the output what the user will get.

INPUT:

Ratings and Reviews from e-commerce website particularly from Flipkart.

PROCESS:

Analyze and categorize review data. The reviews are categorized based as positive and negative considering ratings. 1,2,3 being negative reviews and 4,5 being positive. Finally, the product is recommended based on positive reviews and the positive reviews should be greater than sixty percent of the total reviews given for analyzing.

OUTPUT:

Recommendation of the product based on customer reviews/ratings.

Summary

This chapter summarizes system design and implementation of the system proposed, which is Customer Reviews for Product Recommendation.

CHAPTER: 4

RESULTS AND ANALYSIS

This chapter presents the experimental setup and result analysis carried out for different Test cases

4.1 Experimental Setup

The proposed system is implemented using PYTHON with FLASK under the windows and Linux environment. To test the developed system several products and categories are used.

Users Enters Product URL in the search box. The product is searched from the e-commerce websites and the ratings and reviews are scraped using beautiful soup. Ratings and reviews are categorized and recommended accordingly.

4.2 Search Page

This section contains the Graphical user interface (GUI) implemented.

Figure 4.1 Search page contains two bars where user can enter a product URL and number of reviews to analyze to get a recommendation.

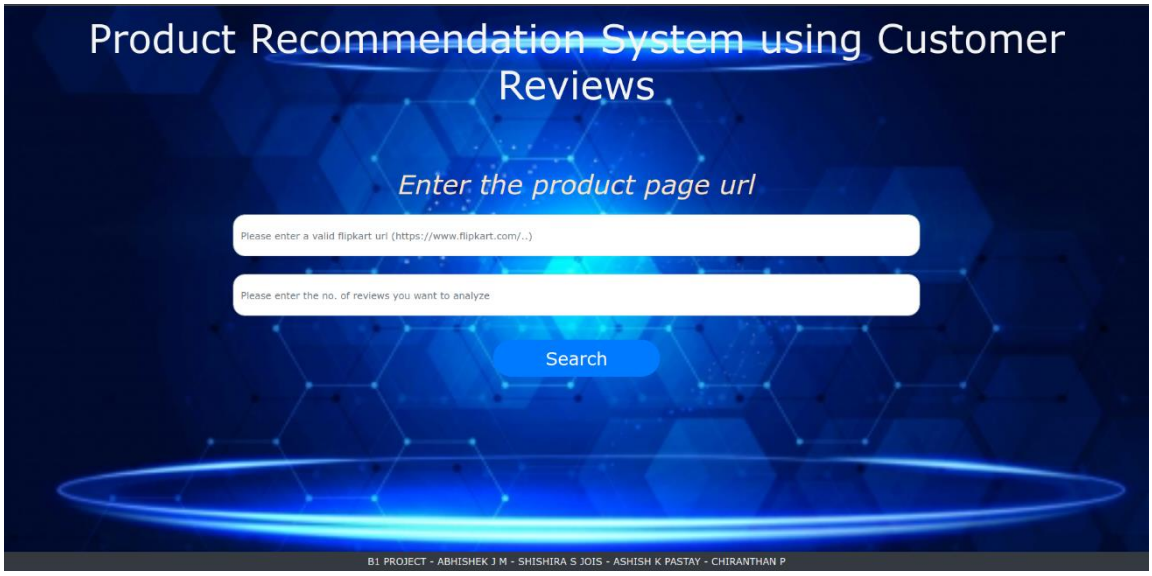

Figure 4.1: Search Page

4.3 Test Cases

This section contains the test cases considered during the implementation of graphical user interface (GUI).

Test Case 01: Empty condition for URL and number of reviews

Figure 4.2 and 4.3 contains empty condition for URL and number of reviews in the home page search bar.

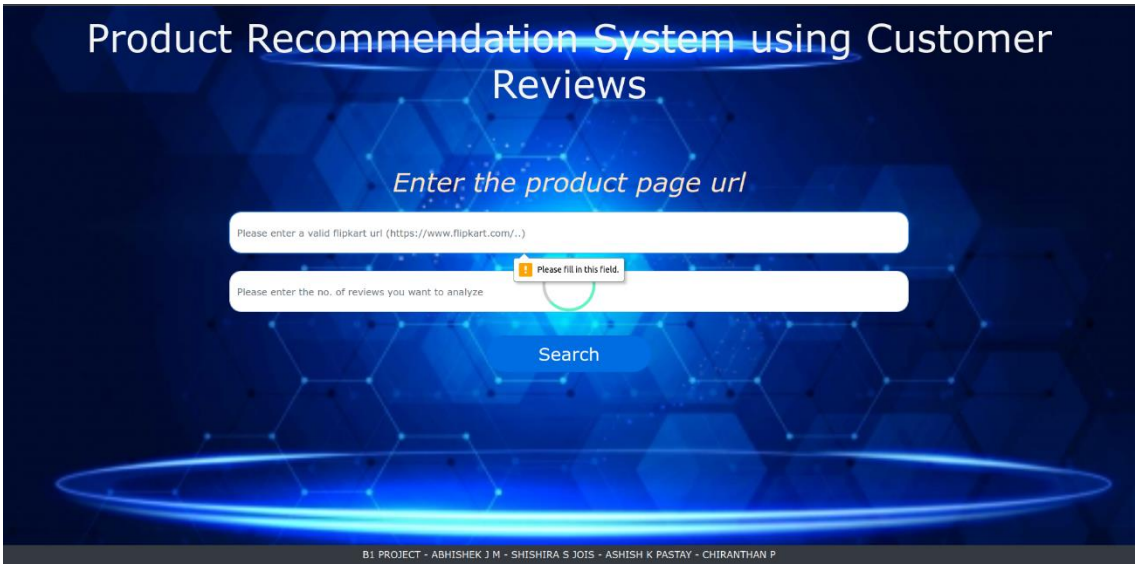

Figure 4.2: Empty condition for URL in search bar

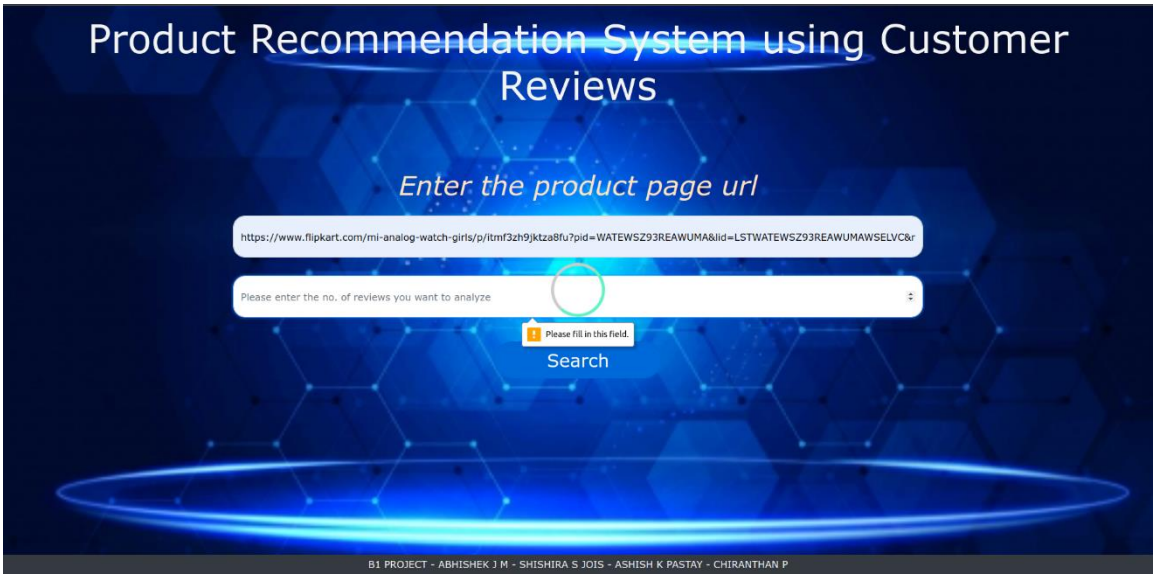

Figure 4.3: Empty condition for number of reviews in search bar

Test Case 02: Invalid URL and Invalid URL format

Figure 4.4 and 4.5 contains Invalid URL and Invalid URL format condition respectively.

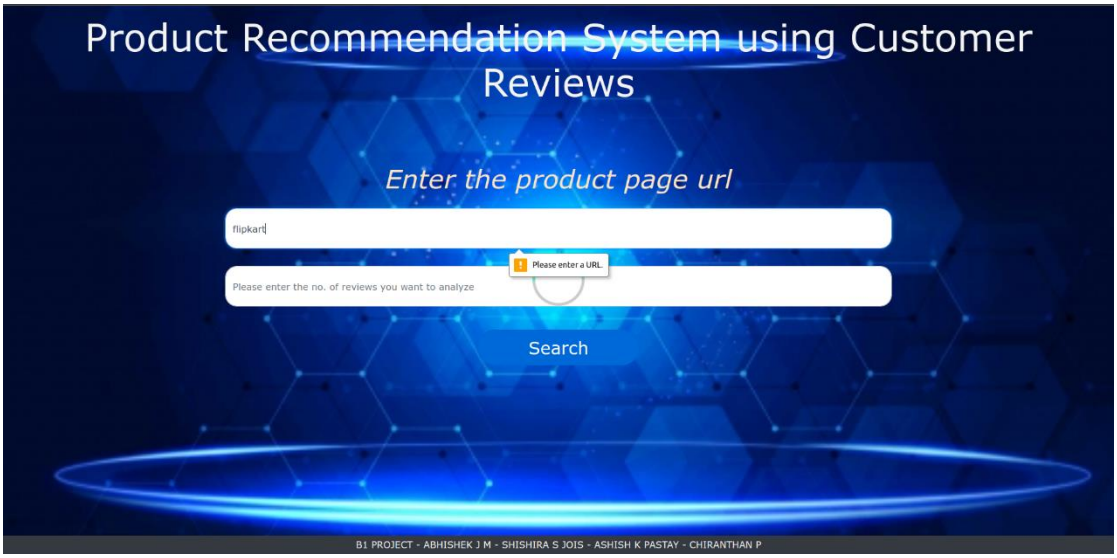

Figure 4.4: Invalid URL

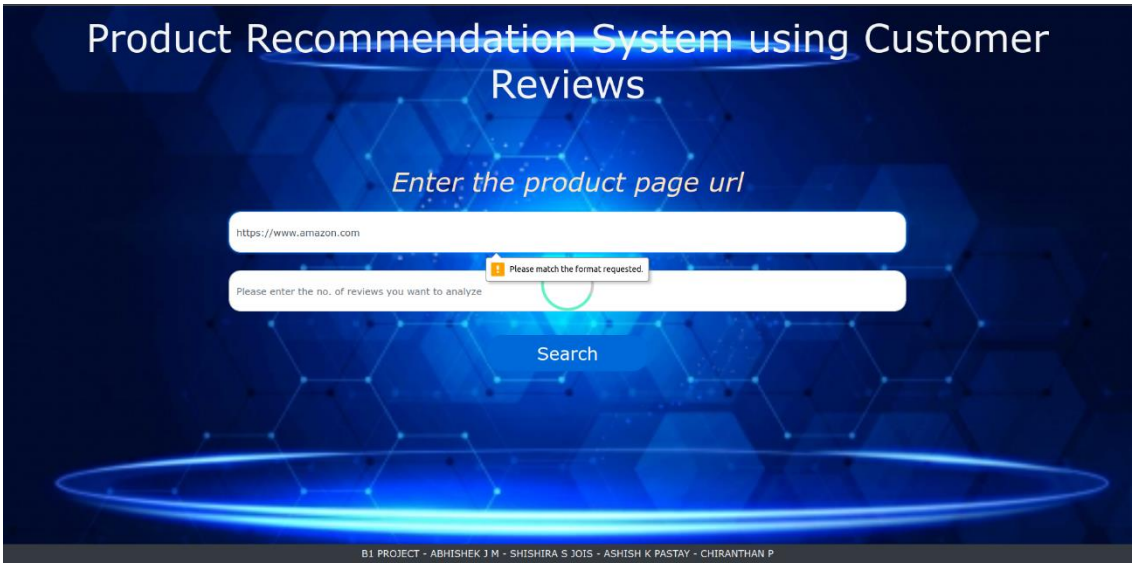

Figure 4.5: Invalid URL format

Test Case 03: Valid URL and Invalid review number

Figure 4.6 contains valid URL with Invalid review number condition.

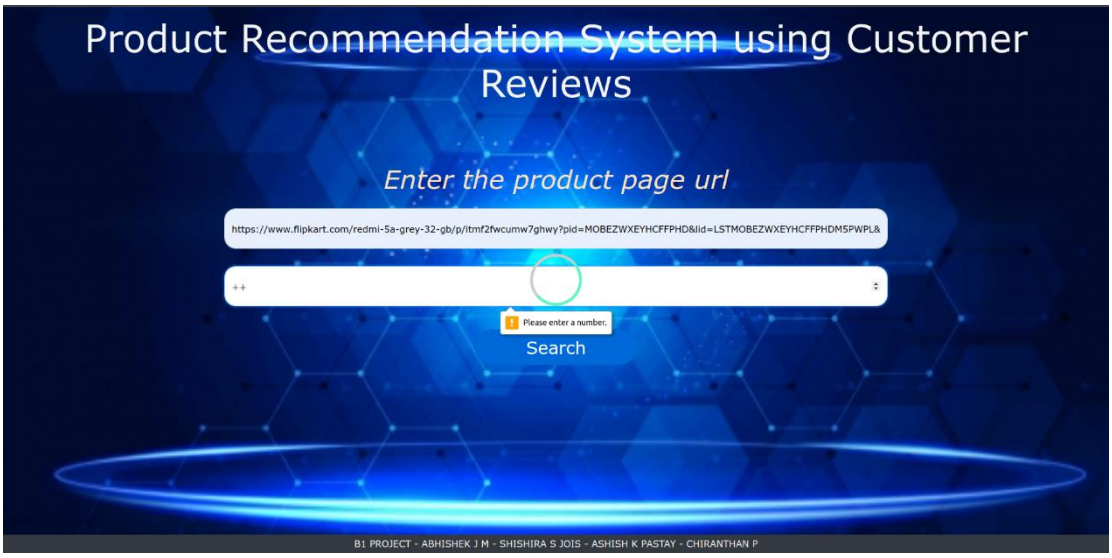

Figure 4.6: Valid URL and Invalid review number

Test Case 04: Valid URL and Invalid review number range

Figure 4.7 contains Valid URL with Invalid review number range condition.

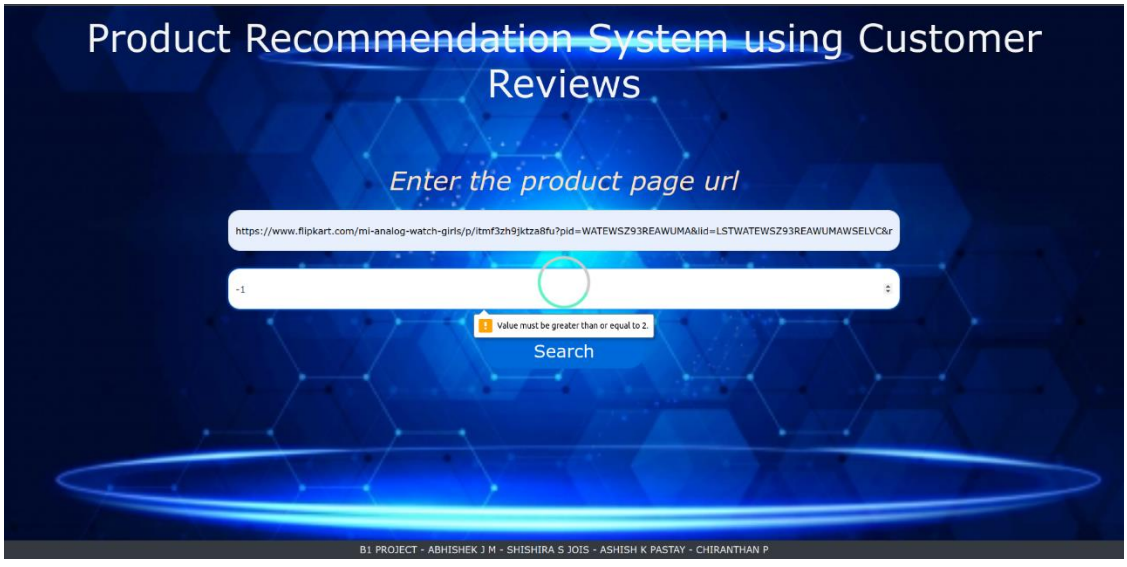

Figure 4.7: Valid URL and Invalid review number range

Test Case 05: Valid URL and Valid number of reviews, Product recommended

Figure 4.8 and 4.9 contains review report which consists of positive and negative reviews categorized based on ratings and recommended accordingly.

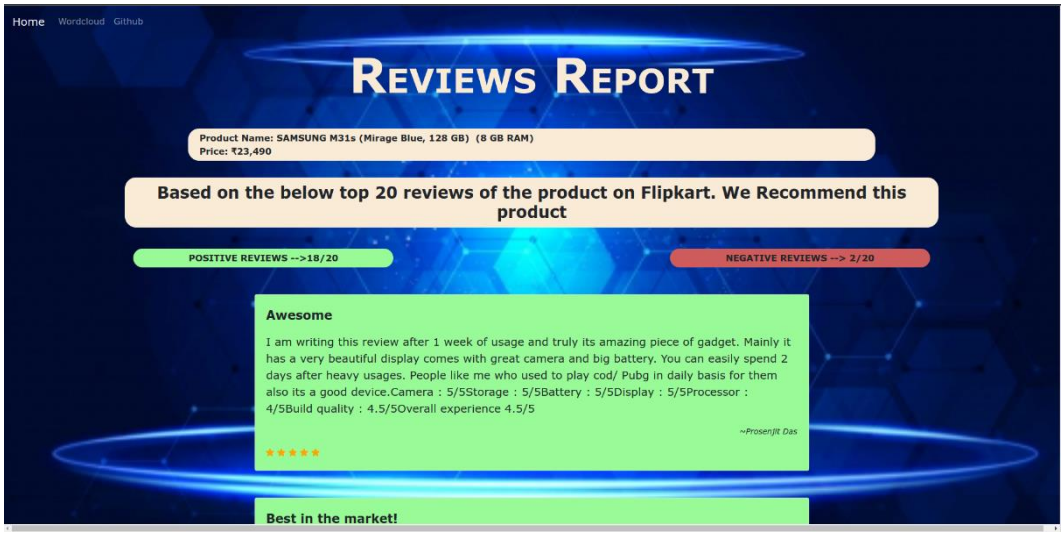

Figure 4.8: Review Report Page 01 (Recommended)

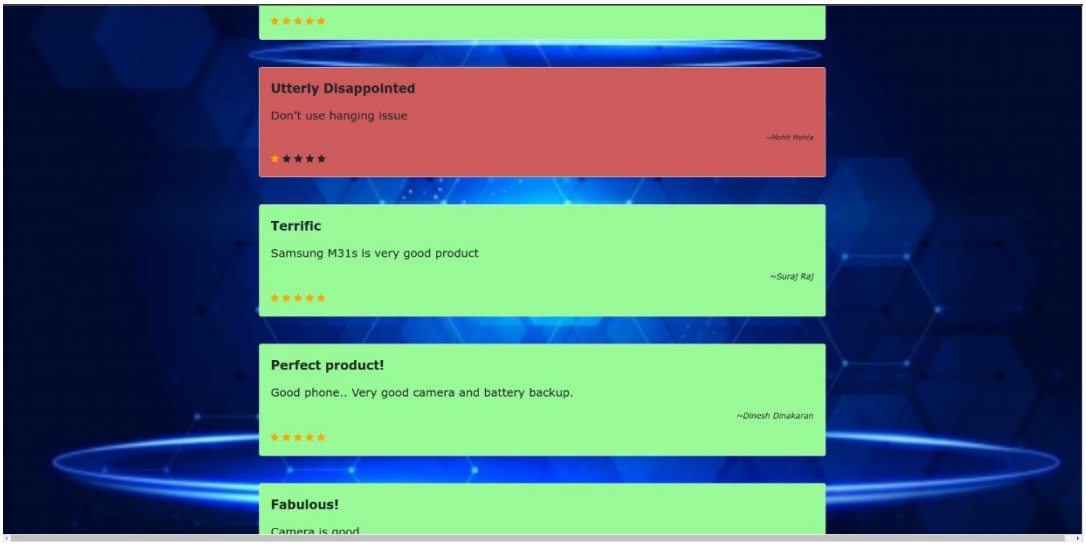

Figure 4.9: Review Report Page 02 (Recommended)

WordCloud Page 01

Figure 4.10 contains textual representation of reviews using Natural Language Processing for recommended product.

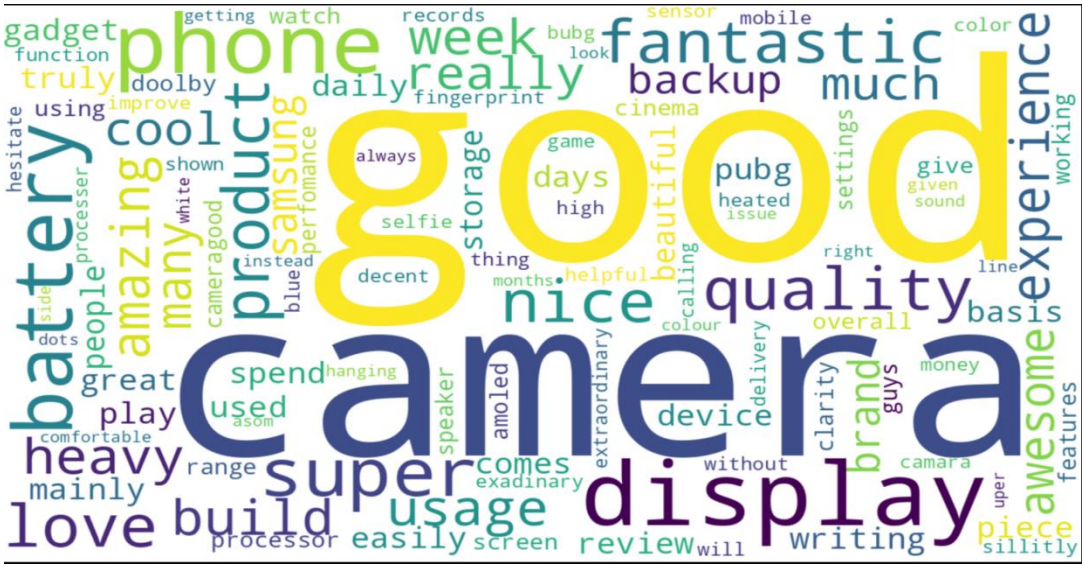

Figure 4.10: WordCloud Page 01

Test Case 06: Valid URL and Valid number of reviews, Product not recommended

Figure 4.11 contains review report which consists of positive and negative reviews categorized based on ratings and the product is not recommended.

Figure 4.11: Review Report Page (Not Recommended)

WordCloud Page 02

Figure 4.12 contains textual representation of reviews using Natural Language Processing for product not recommended.

Figure 4.12: WordCloud Page 02

GitHub Page

Figure 4.13 contains source code of our project.

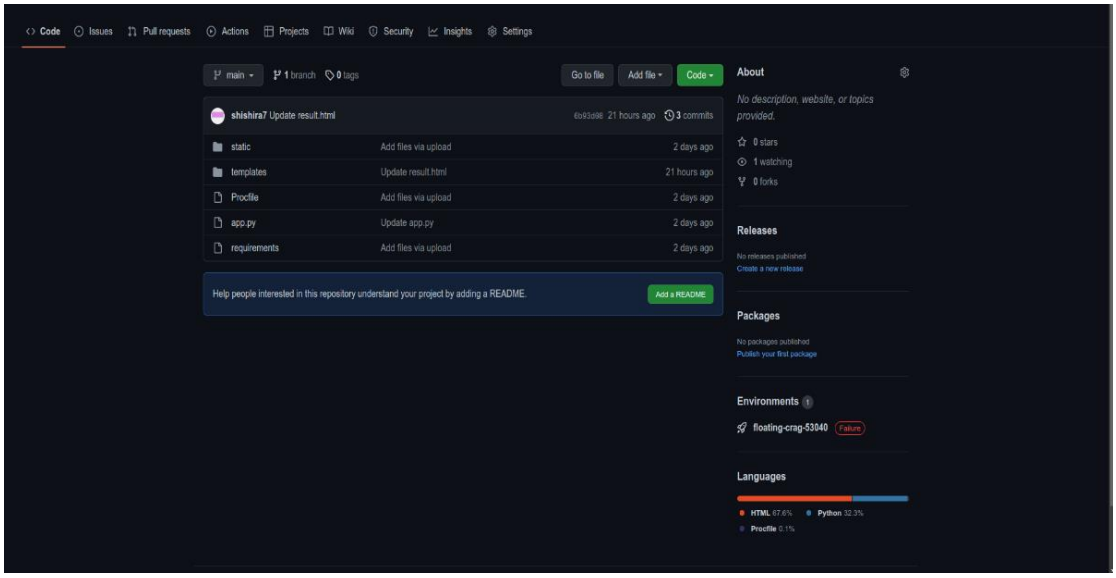

Figure 4.13: GitHub Page

Summary

This chapter summarizes the result and analysis of the proposed system.

CHAPTER: 5

CONCLUSION

An evolutionary shift from offline markets to digital markets has increased the dependency of customers on online reviews to a great extent. Online reviews have become a platform for building trust and influencing consumer buying patterns. With such dependency there is a need to handle such a large volume of reviews and present credible reviews before the consumer's future, the work can be extended to perform multi-class classification of reviews which will provide delineated nature of review to the consumer, hence better judgement of the product. It can also be used to predict the rating of a product from the review. This will provide users with a reliable rating because sometimes the rating received by the product and the sentiment of the review do not provide justice to each other. The proposed extension of work will be very beneficial for the e-commerce industry as it will augment user satisfaction and trust.

FUTURE WORK

Customer have their own priorities in selection of the product. In future work different features of products are considered in the recommendation system. So that customer can buy the products depending on their priorities. The final goal is to deploy a web application supporting a recommendation system**.**

REFERENCES:

[1] Karthik.R.V , Sannasi Ganapathy and Arputharaj Kannan, “A Recommendation System for Online Purchase Using Feature and Product Ranking” in Proceedings of Eleventh International Conference on Contemporary Computing (IC3) 2018.

[2] Jun, M., Yang, Z. and Kim, D. "Customers' perceptions of online retailing service quality and their satisfaction", International Journal of Quality & Reliability Management, 2004.

[3] Pankaj, Prashant Pandey, Muskan and Nitasha Soni, “Sentiment Analysis on Customer Feedback Data: Amazon Product Reviews”, in Proceedings of International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon) 2019.

[4] Stephina Rodney D’souza and Kavita Sonawane, “Sentiment Analysis Based on Multiple Reviews by using Machine learning Approaches”, in Proceedings of the Third International Conference on Computing Methodologies and Communication (ICCMC) 2019.

[5] Jo mackiewicz, Dave yeats, and Thomas thornton, “The Impact of Review Environment on Review Credibility.” in Proceedings ieee transactions on professional communication, vol. 59, no. 2, june 2016.

[6] Thiruni D. Jayasiriwardene, Gamage Upeksha Ganegoda, “Keyword extraction from Tweets using NLP tools for collecting relevant news” International Journal of Computer Applications, vol. 109, no. 2, pp. 18-23, 2015.

[7] Reshma V and Ansamma John, “Aspect Based Summarization of Reviews Using Naïve Bayesian Classifier and Fuzzy Logic”, in Proceedings of the 2nd International Conference on Next Generation Computing Technologies (NGCT-2016) Dehradun, India 14-16 October 2016.

[8] Ujwal B V S , Bharat Gaind , Abhishek Kundu , Anusha Holla and Mukund Rungta., “Classification-based Adaptive Web Scraper”, in Proceedings of the 16th IEEE International Conference on Machine Learning and Applications,2016.

[9] Tanya Gera, Deepak Thakur and Jaiteg Singh. "Identifying Deceptive Reviews Using Networking Parameters", in proceedings with IEEE International conference on networking parameters,9-11, 2015.

[10] Jawad khan, byeong soo jeong, "Summarizing Customer Review based on Product Feature and Opinion", in Proceedings with 2016 International Conference on Machine Learning and Cybernetics, Jeju, South Korea, 10-13 July, 2016.

[11] Mohammad F. A. Bashri, Retno Kusumaningrum, "Sentiment Analysis Using Latent Dirichlet Allocation and Topic Polarity Wordcloud Visualization" in proceedings with Fifth International Conference on Information and Communication Technology (ICoICT), 2017.