

Informatics Institute of Technology
In collaboration with
University of Westminster, UK.

Facebook Based E-Commerce Product Rating Using
Opinion Mining.

A dissertation by
Hetti Arachchige Maneendra Madushani Perera

Supervised by
Guhanathan Poravi

Submitted in partial fulfillment of the requirements for the
M.Sc. in Advanced Software Engineering
Department of Computing

August 2017

©The copyright for this project and all its associated products resides with Informatics
Institute of Technology

Declaration

I hereby certify that this project report and all the artifacts associated with it is my **own** work and it has not been submitted before nor is currently being submitted for any degree programme.

Full Name of Student: Hetti Arachchige Maneendra Madushani Perera

Registration No: 2013130

Signature:Maneendra Perera.....

Date:01 – 08- 2017.....

Abstract

A business using social media like Facebook to launch products and strengthen the existing products is becoming an emerging trend during the recent past. In addition, buyers are tending to get the buying decisions based on the comments or reviews posted in social media. Due to having large no of comments or reviews (opinions) in the web, it has been difficult to get a clear idea whether the product is good or bad, what are the features of the product and extract the opinions manually.

The focus of this research is to design and develop an opinion mining system, which is capable of mining the Facebook post comments or reviews, and present a summarization on the features of the products and their polarity automatically. In addition, the system should be able to categorize the comments or reviews as positive or negative and rate them from 1 – 5. System should provide a summary on sentence level and feature level but providing a document level summarization is out of scope of this project.

Automatic opinion mining can be done in three levels document level, sentence level and aspect level. In the project both sentence level and aspect level opinion mining are considered and to opinion mining process syntactic based approach has been used. Syntactic dependency, aggregate score of opinion words, SentiWordNet score and aspect table are used in the process. Apriori algorithm is used to identify the frequent aspects in the reviews, which are mostly considered in the system. Finally, two summarization tables – aspect table and sentence table are created and presented to the user. The developed system is tested using a data set on camera reviews, which was taken from the web and preprocessed before using it for testing.

The developed system will be useful for vide variety of people like business owners, buyers, marketing personal, competitors etc. to get a better understanding on the products they sell, product they planning to buy or product that they are competing with

Keywords: Opinion Mining, Sentence Level, Aspect Level, SentiWordNet, Apriori Algorithm, Aggregate Score

Acknowledgment

Firstly, I would like to express my gratitude to my supervisor Guhanathan Poravi for the support and encouragement given throughout the project. I would not have been successful without his guidance and will not be able to finish this project without his supervision and motivation.

Furthermore, I would like to thank my mother, husband and my friend Naveen who helped me in many ways to make this project a reality.

In addition, I would like to thank the participants in my evaluation survey who have willingly share their precious time, ideas and suggestions during the evaluation.

I will be grateful for all the support and love you have given me during this rush period of my life.

Table of Contents

List of Figures	7
List of Tables	9
Acronyms.....	10
1. Introduction	11
1.1. Introduction	12
1.2. Background Study.....	12
1.3. Problem Domain	14
1.4. Title	15
1.5. Aims and Objectives	15
1.5.1. Aim	15
1.5.2. Aim in Detail.....	15
1.5.3. Objectives	15
1.6. Significance of the Study	16
1.7. Scope of the Study.....	17
1.8. Limitations of the Study.....	17
1.9. Summary	18
2. Literature Survey	19
2.1. Introduction	20
2.2. Challenges	21
2.3. Social Media.....	23
2.4. Data Mining.....	24
2.5. Opinion Mining or Sentiment Analysis	25
2.6. Opinion Mining Approaches.....	28
2.7. Opinion Mining Levels	30
2.8. Opinion Mining Process.....	31
2.9. Opinion Summarization	34
2.10. Opinion mining tools	34
2.11. Study of Past Research	34
2.12. Summary.....	46
3. Research Methodology	47
3.1. Introduction	48
3.2. Research Methodology.....	48
3.3. Development Methodology.....	48
3.4. Project Management Methodology	49
3.4.1. Time Management	50
3.4.2. Risk Management	50

3.5. Summary	51
4. Requirement Specification	52
4.1. Introduction	53
4.2. Rich Picture	53
4.3. Context Diagram	54
4.4. Analysis	54
4.4.1. Stakeholder Analysis	54
4.4.2. Requirement Analysis	55
4.4.3. Use Case Model	57
4.4.4. Use Case Description	58
4.5. Summary	60
5. System Design and Architecture	61
5.1 Introduction	62
5.2 Proposed OM System	63
5.3 High Level Design of the Opinion Mining System	64
5.4. System Architecture	64
5.5. Data Flow Diagram	65
5.6. Class Diagram	66
5.7. Sequence Diagrams	68
5.8. Summary	72
6. Implementation	73
6.1. Introduction	74
6.2. Comments Retrieval	74
6.3. Pre Processing	77
6.3.1. Remove Duplicate Author Comments	77
6.3.2. Remove Question Based Comments	77
6.3.3. Case Normalization	78
6.3.4. Comments Tokenization	78
6.3.5. Remove Stop Words	78
6.3.6. Stemming	80
6.3.7. POS Tagging	80
6.3.8. Aspect / Feature Extraction	83
6.3.9. Apriori Algorithm	83
6.4. Processing	84
6.4.1. Remove Objective Comments	84
6.4.2. Opinion Words Identification	85
6.4.3. Calculating Scores	90

6.5.	Product Rating.....	94
6.6.	UI Implementation	95
6.7.	Deployment of the Application	97
6.8.	User Manual	97
6.9.	Summary	97
7.	Testing	98
7.1.	Introduction	99
7.2.	Testing Process.....	99
7.2.1.	Planning and control	99
7.2.2.	Analysis and design	100
7.2.3.	Implementation and Execution	101
7.2.4.	Evaluating exit criteria and reporting	104
7.2.5.	Test Closure Activities.....	106
7.3.	Limitations of the testing process.....	106
7.4.	Summary	107
8.	Evaluation.....	108
8.1.	Introduction	109
8.2.	Evaluation Methodology	109
8.3.	Selection of Evaluators.....	109
8.4.	Evaluation Findings.....	109
8.5.	Self-Evaluation.....	110
8.6.	Summary	111
9.	Conclusion	112
9.1.	Introduction	113
9.2.	Discussion and Conclusion	113
9.3.	Contribution	114
9.4.	Learning Outcomes	114
9.5.	Recommendation and Future Work	115
9.6.	Concluding Remarks	117
9.7.	Summary	117
	References	118
	Appendices.....	121
	Appendix A: GUI Validation Test Figures	121
	Appendix B: User Manual.....	121
	Appendix C: Gantt chart	125
	Appendix D: Questionnaire Used for Evaluation	126

List of Figures

Figure 1: Social Media Sites

Figure 2: Hierarchy of Data Mining

Figure 3: Types of Opinions

Figure 4: OM Categories Based on the Method Used in Mining

Figure 5: OM Approaches

Figure 6: Proposed System Architecture

Figure 7: Working of Proposed System Architecture

Figure 8: Opinion Tree

Figure 9: Feature Categories

Figure 10: Research Methodology

Figure 11: Rich Picture of OM Problem

Figure 12: Context Diagram of OM Application

Figure 13: Power /Interest Grid for Stakeholder Prioritization

Figure 14: Use Case Diagram

Figure 15: System Architecture and Design

Figure 16: Opinion Mining System Overview

Figure 17: High Level Design of Opinion Mining System

Figure 18: System Architecture

Figure 19: Data Flow Diagram

Figure 20: Class Diagram for System

Figure 21: Class Diagram for System Loading

Figure 22: Sequence Diagram for Login

Figure 23: Sequence Diagram for Retrieve Facebook Comments

Figure 24: Sequence Diagram for Pre Process Comments

Figure 25: Sequence Diagram for Process Comments

Figure 26: Sequence Diagram for Find Relations and Calculate Scores

Figure 27: Facebook User Access Token

Figure 28: JSON Object and Array

Figure 29: Facebook App Settings

Figure 30: Tokenization Sample

Figure 31: POS Example

Figure 32: Basic Dependencies

Figure 33: Enhanced Dependencies

Figure 34: nsubj Dependency
Figure 35: amod Dependency
Figure 36: advmod Dependency
Figure 37: conj Dependency
Figure 38: cop Dependency
Figure 39: acomp Dependency
Figure 40: advcl Dependency
Figure 41: dobj Dependency
Figure 42: neg Dependency
Figure 43: Weighted Average
Figure 44: SentiWordNet
Figure 45: Aggregate Adverb Score Calculation Logic
Figure 46: Adjective/Verb Score Calculation Logic
Figure 47: Aspect / Sentence Rating Logic
Figure 48: Login Page
Figure 49: Home Page
Figure 50: Aspect Result Table
Figure 51: Sentence Result Table and Summary
Figure 52: Configuration File Content
Figure 53: Validation and Verification in Testing
Figure 54: Aspect Identification Analysis
Figure 55: Aspect Polarity Identification Analysis
Figure 56: Aspect Result Summary Analysis
Figure 57: Positive Comments Analysis
Figure 58: Negative Comments Analysis
Figure 59: Comments Result Summary Analysis
Figure 60: Aspects List Box

List of Tables

Table 1: UC-01: Login System

Table 2: UC-02: Submit Post Id

Table 3: UC-03: View Results

Table 4: UC-04: Logout System

Table 5: UC-05: Send Post Comments

Table 6: REST Operations

Table 7: Stop Words List

Table 8: POS Tags

Table 9: Functional Test Cases

Table 10: Non Functional Test Cases

Table 11: Issues and Solutions Table

Acronyms

Abbreviation	Definition
OM	Opinion Mining
SVM	Support Vector Machine
POS	Parts of Speech
acomp	adjectival complement
advcl	adverbial clause modifier
advmod	adverb modifier
amod	adjectival modifier
appos	appositional modifier
auxpass	passive auxiliary
cc	coordination
ccomp	clausal complement
conj	conjunct
cop	copula
REST	Representational State Transfer
HTTP	Hypertext Transfer Protocol

1. Introduction

Problem Domain

Title

Aims and Objectives

Significance of the Study

Scope of the Study

Limitations of the Study

Summary

1.1. Introduction

The first chapter will provide details about background of the project, problem domain, aims and objectives of the project. In addition, it will elaborate the significance of the study while describing the scope and limitations of the study.

1.2. Background Study

Internet changed the way of doing business and introduced a new business model Electronic Commerce that is well known as e-commerce. The history of e-commerce goes to 1970s and it became one of the most popular business models around the world by 1990s with the evolution of eBay and Amazon.

Social networking (social networking site, social media, social networking service), online platform which enables people to share their ideas, interests, activities etc. and build up social relations. Social networking has been a popular topic in the recent past and variety of social networking services like Facebook, LinkedIn are available online. Facebook is the most popular social networking site around the world.

Social media can be categorized into three main categories according to their primary objective,

1. Social networking services for socializing with existing friends(Facebook)
2. Social networking services for nonsocial interpersonal communication(LinkedIn)
3. Social networking services for helping users to find specific information or resources(Goodreads)

According to a study by John S. and James L. Knight Foundation (2015), 63% people who use Facebook and Twitter consider those two networks as their main source of news and entertainment news. Also as a study by Mody Milind (2015), 85% people in age group 18 to 34 use social networking sites for decision making when purchasing a good or service while over 65% people in age group 55 and over rely on word of mouth. These two figures clearly show the changes of thinking pattern in the society and how powerful the social media in modern society.

When consider about the social media and business, it seems to have a major impact on the businesses around the world by connecting people at a lower cost.

Facebook

Facebook was launched on 4th of February 2004 and now it has more than 1.65 billion active users as of March 31st 2016. Facebook's main revenue comes from advertisements. Even though the primary objective of Facebook is to make a network for socializing with friends, it wants to be more on ecommerce by being the next impulse buy. Facebook allows the businesses to display products on the pages and let customers to checkout and pay while in the page.

In 2009 companies, started selling products directly via Facebook and in 2011 top brands like GameStop, Gap, and J.C. Penney started doing businesses on Facebook.

Facebook Page

As Facebook, a Facebook page is a free public profile specifically created for businesses, brands, celebrities, causes, and other organizations and provides a voice on Facebook. It is visible to everyone by default and any person in Facebook can connect to these pages by becoming a fan. Facebook pages give an opportunity for a business to market the business and introduce products services to the customers. It also allows promoting the items via News Feed, which can be seen by the customers. Facebook pages are easy to setup and handle. Following are some of the features of pages,

- Facebook pages work great on mobile devices and it helps customers to easily communicate with the business.
- Facebook pages are designed in a way that customers find easy to learn about the business, products and services.
- Facebook pages provide an insight about how people found the Page, visitor demographics such as age, gender, location etc. that helps to identify the audience properly, promote the business effectively and keep growing the business.

Doing businesses via Facebook is becoming more and more popular and therefore it has been essential to identify the strategies to strengthen the business while increasing the profits.

1.3. Problem Domain

E commerce has been a popular topic over the years around the world and many users purchase products through E-commerce websites. With the rise of social networks, there was a huge impact on e commerce industry and the number of users who engaged in the industry increased tremendously. According to a study conducted by leading e commerce vendor Kaymu, in Sri Lanka it is expected that there will be 72% growth in e-commerce transactions in the near future. It is a good e commerce strategy to analyze the comments of users, study the behaviors of users, identify the trends and then craft effective strategies forecasting the future needs. As described in paper by Alexandra Cernian *et al.* (2015) according to KRC research 65% of electronic devices are influenced by reviews when selecting brands and products. Also 77% of people are interested in other consumer reviews.

Opinion mining or semantic analysis is a type of natural language processing for tracking the mood or views of the public about the specific product. It is useful in several ways. For a company it is very important to get feedback from clients or prospective clients (Alexandra Cernian *et al.*, 2015). For marketers it will be helpful to evaluate the success of new product launch, identify the most popular products, and measure the customer satisfaction and the reasons for them. Then they can enhance and improve the products based on the opinions of customers. For buyers it will be helpful to identify the good products, build a trust and based on that make decisions whether to buy the product or not. When a buyer wants to buy a product from internet, he is interested in knowing other's opinion about the product. Then based on the opinion his decision making process will be continued.

In modern days' Social network based E commerce has been a very popular method and putting comments about the products and customer review posts are becoming popular among the most customers. However, the issue is since there are huge number of comments and posts, it has been difficult to get a clear idea about the product, whether the comments or posts are positive or negative and for how extent. Therefore, opinion mining application would be a great idea in this scenario, which evaluates the opinions, and rate them accordingly and now opinion mining trend is moving to the sentimental reviews of twitter data, comments used in Facebook on pictures, videos or Facebook status (G. Angulakshmi *et al.*, 2014). In addition, it has become the current trend to create automatic tools to process consumer reviews and opinions about products or services. Then it will be a great help for the marketers as well as buyers to identify the product quality, product popularity etc.

1.4. Title

Facebook based e commerce product rating using opinion mining.

1.5. Aims and Objectives

Lot of research solutions have been implemented on E commerce product rating but Facebook based ecommerce product rating will be a novel area to be researched. In addition, there is a huge improvement for E commerce product rating because of the complexity of the area. Every current solution has their own limitations that need to be addressed further. Therefore, aim and objective of this system is to minimize the existing limitations and develop an effective Facebook based E commerce product rating application, which helps the E commerce community to their development.

1.5.1. Aim

Design and develop an effective application, which can effectively and efficiently rate the products based on customer's opinion mining.

1.5.2. Aim in Detail

Even though there are several solutions in this regard, they have their own limitations. Therefore, aim of this project is to design and develop an effective system, which is capable of mining opinions of the customers, and rate the products accordingly. Main challenge would be to identify an effective, efficient and suitable opinion mining methodology, algorithm. Extensive research will need to be carried out in the area and the most suitable methodology will be adopted in the application. Results will be presented in an attractive and user-friendly manner to the non-technical users.

1.5.3. Objectives

- Prepare the project proposal
- Carry out a literature review reading the current researches and evaluating the current solutions
- Carry out a study on current solutions and their limitations

- Carry out a study on currently available opinion mining tools, their usage and how to use them in an application
- Carry out a study on the areas that need to be improved
- Identify the best opinion mining methodology, algorithms, processes, tools etc. and select one of from each area
- Prepare the requirement specification
- Carry out the general architectures used in currently available solutions and design the system
- Build a prototype of the system
- Develop the system
- Critically evaluate the system by conducting testing and evaluation surveys

1.6. Significance of the Study

Social network based e commerce has become popular in the recent past and most of the people tend to have a high interest on them. Also lot of researches have proven that comments or reviews in social media sites about a product or service does a major impact on decision making among young generation (Alexandra Cernian *et al.* ,2015). However, due to large number of information in the social media sites it has become difficult to get and clear understating manually. This project comes to play in this point and present a way of automatically discover valuable knowledge from the opinions stated and present it to the user in a user friendly way. Opinion mining plays a vital role for decision support system for marketers since the people provide a valuable suggestions and opinion on their products and services (M. Lovelin Ponn Felciah and R.Anbuselvi, 2015), (Alexandra Cernian *et al.*, 2015).

Opinion mining can be categories in to three main levels document level, sentence level and aspect level. In this study two categories, sentence level and aspect level opinion mining are researched and analyzed.

Furthermore, this study included the Facebook integration and it will be helpful for the future as Facebook is now focusing on businesses unlike past and they are keen popularizing Facebook based E commerce.

1.7. Scope of the Study

In this research only sentence level and aspect level opinion mining is covered for direct opinions. Providing a document level opinion mining summarization is out of the scope in this phase and it will be considered in the next phase. In a product or service, different users will be interested in different aspects. E.g: - Two persons are trying to buy cameras and reading reviews. Person A wanted to find a camera with least cost while other person wants to buy a camera with good quality. Therefore, without knowing the interested aspect it would not be effective to provide a document level summary, whether the product is positive or negative. In the next phase, the application will allow users to select their interested aspects and then the application could provide document level result whether the product or service is positive or negative by going through all the reviews against the aspects selected. In addition, the study does not consider the domain specific words and for opinion mining mix of syntactic based approach and resource based approach using SentiWordNet will be used. In the study only English language opinion mining is analyzed.

1.8. Limitations of the Study

In this research, following areas will not be covered and they will be considered as future work.

- Sarcasm detection

Sarcasm is a pervasive linguistic phenomenon in online documents that express subjective and deeply felt opinions. Sarcasm reverses the sentiment polarity of the literal sentiment (Rajeev Arora et al., 2014). Sarcasm detection is a challenge in OM. In this study, identifying sarcasms are not covered.

- Identify opinion spamming

Opinion spamming is illegal activities like writing fake reviews, writing underserving positive comments to promote products, writing negative comments to damage reputation of products etc. with the intention of misleading readers. Detecting such scenarios are becoming more important and critical. However, in the study opinion spamming is not studied and it will definitely come in future work section. Opinion spamming is also a unique challenge in OM (Rajeev Arora et al., 2014).

- Authors behavioral identification – identify authors motivation

This is similar as identify opinion spamming. People tend to post opinions based on their personal agendas like writing positive comments to help business of their friends,

writing negative comments to damage the business of their competitors or rivals. Such situations are not analyzed in this study.

- Identify implicit aspects

Implicit aspects are which are not directly mentioned or expressed in the comment. Implicit aspect extraction is a complex problem (Rajeev Aurora and Srinath Sirinivasa, 2014). Such aspects are not covered in this study.

- Correcting spelling mistakes

- Constructive opinion identification

Most of the constructive opinion sentences are written in future tense and use “would be”, “could be” etc. Constructive opinions need not to imply that the opinion holder is negatively inclined about the entity (Rajeev Arora et al., 2014). Identifying constructive opinions will be covered in next phase.

1.9. Summary

In this chapter, a detail description has been given on the background of the study and problem domain. Then the aim and objectives of the project are presented. Finally, a brief description is given on the significance of the study while describing the scope and limitations.

2. Literature Survey

Introduction	
Challenges	
Socail Media	
Data Mining	
Opinion Mining	
Opinion Mining Approaches	
Opinion Mining Levels	
Opinion Mining Process	
Opinion Summarization	
Opinion Mining Tools	
Study of Past Research	
Critical Evaluation of Literature	
Summary	

2.1. Introduction

Ideas or opinions of others always influence our day-to-day life. We always feel comfortable if we get positive comments on the product or service that we are planning to buy or consume. Various groups keep an eye on the opinions like marketing professionals to evaluate the products and services released to the market, identify the prospective markets, identify the niche markets etc. Consumers to get feedback on the product or service and make decisions on buying, product designers to identify the weaknesses and improve the products, identify the requirements of the consumers. Researches to build innovative products or services and introduce them in to the relevant markets

With the rapid evolution of web, blogging, social networking, twitting etc. which are called as User Generated Content (Rajeev Aurora and Srinath Sirinivasa, 2014) have increased immensely and people have used to share their ideas, views, opinions via them (Amani K Samha, 2016). Opinion mining has become a growing and interesting area of research in natural language processing and information retrieval in past few years (Onifade O.F.W and Malik M.A., 2015). The ultimate aim of the researches is to present information effectively and easily. Opinion mining or Sentiment analysis refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. It extracts people's opinions and analyzes people's opinions, appraisals, attitudes and emotions towards organizations, entities, persons, issues, actions, topics and their attributes. As Vijay B. Raut and D.D. Londhe (2014), opinion mining is classification of user's expressed opinion in to positive or negative polarity.

As Onifade O.F.W and Malik M.A. (2015) with the fastest growing web, it is very difficult for a user to read and understand the content. Therefore, it is important to identify and extract important information using opinion mining, which is a summarization process. Opinion mining is one of the most popular research topic and several applications have been built on decision-making, recommendation systems, feedback analysis etc. Today sentiment analysis or opinion mining plays a vital role for decision support system for the marketers since the people provide a valuable suggestions and opinion on their products and services (M. Lovelin Ponn Felciah and R. Anbuselvi, 2015). In the past most of the opinion mining work done has done in document level or sentence level and it did not help to discover what exactly people liked or did not like (A. Jeyapriya et al., 2015). It only classifies a document/text to be "subjective" or "objective" which has no practical significance. Therefore, now most of the researches are moving towards phrase level opinion mining

which performs fine-grained analysis and directly looks at the opinion (A. Jeyapriya et al., 2015). It classifies the opinion as “positive”, “negative” or “neutral”.

Opinion mining is a hard problem to be solved due to the highly unstructured nature of natural language and difficulty of a machine to interpret the meaning of a sentence (Chinsha T C and Shibily Joseph, 2015). As Amani k Samha (2016) although the process seems straightforward, it involves huge amount of work due to the complexity of natural language and the huge number of reviews.

Firstly, a broad analysis is done on the existing work carried on the area. Then the focus is on various algorithms developed, approaches to mine the opinions and their strengths and weaknesses. Then a detail analysis is done on the methodologies or processes that uses the best opinion mining algorithms developed. Finally, the focus is on the tools, basically the open source tools that are built for opinion mining and opinion summarization. Apart from those, a report is included with the various challenges and issues faced in the arena. The goal of this research is to find out the opinion mining tools that are capable of integrating in to the proposed system, which extracts opinions from the Facebook.

2.2. Challenges

1. Handling negation words like ‘not’, ‘don’t’, ‘never’ when calculating polarity of a text. Negation should be handles appropriately to get the contextual information of a sentence. (Chinsha T C and Shibily Joseph, 2015)
2. User written review are highly unstructured and there are spelling mistakes. Therefore, it will not get correct syntactic dependency. (Chinsha T C and Shibily Joseph, 2015)
3. Sentiments or opinions are expressed differently in different domains and it is expensive to interpret data from each novel domain (A. Jeyapriya et al., 2015). Word polarity is domain specific. Quiet is a positive sentiment for a car while it is a negative sentiment for a phone (Rajeev Aurora and Srinath Sirinivasa, 2014).
E.g: - Heat in a low time is a positive aspect when talks about a burner but the same is a bad aspect when talks about a phone. Therefore, it is important to classify the opinion considering the domain. One opinion that is positive in a domain can be a negative opinion in another domain.

Polarity is also context specific with the domain. Long is positive sentiment for battery life while it is a negative sentiment for startup time for of the laptop.

4. Document level and sentence level opinion mining information is not sufficient for valuable decision-making. Therefore, phrase level opinion mining is essential for future opinion mining tools or systems (A. Jeyapriya et al., 2015).

E.g: - If a reviewer posted a positive review, it does not mean that he likes all the aspects of a given item. He may like several aspects of the item and may not like several on it.

5. Identify the implicit aspects. Implicit aspects are not explicitly mentioned in a sentence but implied. (Chinsha T C and Shibily Joseph, 2015).

E.g: -This camera is so expensive. Here “price” is not mentioned in the sentence, but implied. Identifying this type of implicit aspect is a great challenge in opinion mining.

6. Understanding domain specific opinion words and their polarity (Rajeev Aurora and Srinath Sirinivasa, 2014).

7. Identify language specific opinion rules (Rajeev Aurora and Srinath Sirinivasa, 2014).

8. Opinions are subjective in nature, and the trust and credibility assigned to an opinion depends on who is giving the opinion and what is their motivation in publically stating their opinion (Rajeev Aurora and Srinath Sirinivasa, 2014).

E.g.: -Trust level of an opinion from a stranger and a friend or relative has a difference.

9. Identifying the constructive opinions, which are suggestions or improvements to improve or enhance the product or the service. It should not be implied as a negative comment about the entity. (Rajeev Aurora and Srinath Sirinivasa, 2014).

10. Identify opinion spamming scenarios. Business may flood social media with positive sentiments about their products and services, which will motivate the prospective customers in an unethical way. Also in tweets, it is noticed that high percentage are promotional tweets and ads, which do not offer any opinion (Onifade O.F.W and Malik M.A., 2015). These kind of scenarios are hard to identify and challenge the opinion mining process.

11. Lexicon, which are used in identifying the opinion structure, is language specific. Lexicons for English language and for Chinese language has a huge difference.
12. Sarcasm detection. Sarcasm reverses the sentiment polarity of the literal sentiment expressed in the document and it is even people find it difficult to recognize. (Rajeev Aurora and Srinath Sirinivasa, 2014), (Chinsha T C and Shibily Joseph, 2015). To better accuracy in sentiment polarity sarcastic statements should be predict to a certain extent (Onifade O.F.W and Malik M.A., 2015).

E.g.: -The country's economy is in an excellent state; it can only go up.

2.3. Social Media

Social media are computer-mediated technologies that allow the creating and sharing of information, ideas, career interests and other forms of expression via virtual communities and networks. Social media is emerging rapidly in recent years and it has become a decision making factor in day-to-day lives of people. Social media created the opportunity for the people to voice their opinions publically.



Figure 1: Social Media Sites

Social media use both web based and mobile technologies to create highly interactive environment to which individual, communities and organizations can share, discuss the ideas and contents. Social media has unique qualities like quality, reach frequency, usability etc. that are different from traditional broadcasting media like TV, radio and papers. Above diagram shows the icons of most

popular social media websites, Facebook, YouTube, Twitter, Skype, Instagram, LinkedIn, Google+ etc. When consider the impact of social media it has been observed that there are both positive and negative impact to the society like stay connected with everyone, help to effective communication in marketing, depression by heavy usage etc.

2.4. Data Mining

Data mining is the process of identify patterns, rules from large volume of data and it is one of the phases in knowledge discovery process. Gain valuable knowledge is the main objective of data mining. Key objectives of the data mining process are to effectively handle large scale of data, mine actionable rules, patterns and gain insightful knowledge (A. Jeyapriya, 2015). Artificial intelligence, machine learning, statistics etc. are used in the data mining process. Opinion mining or Sentiment analysis is a sub disciplinary of data mining.

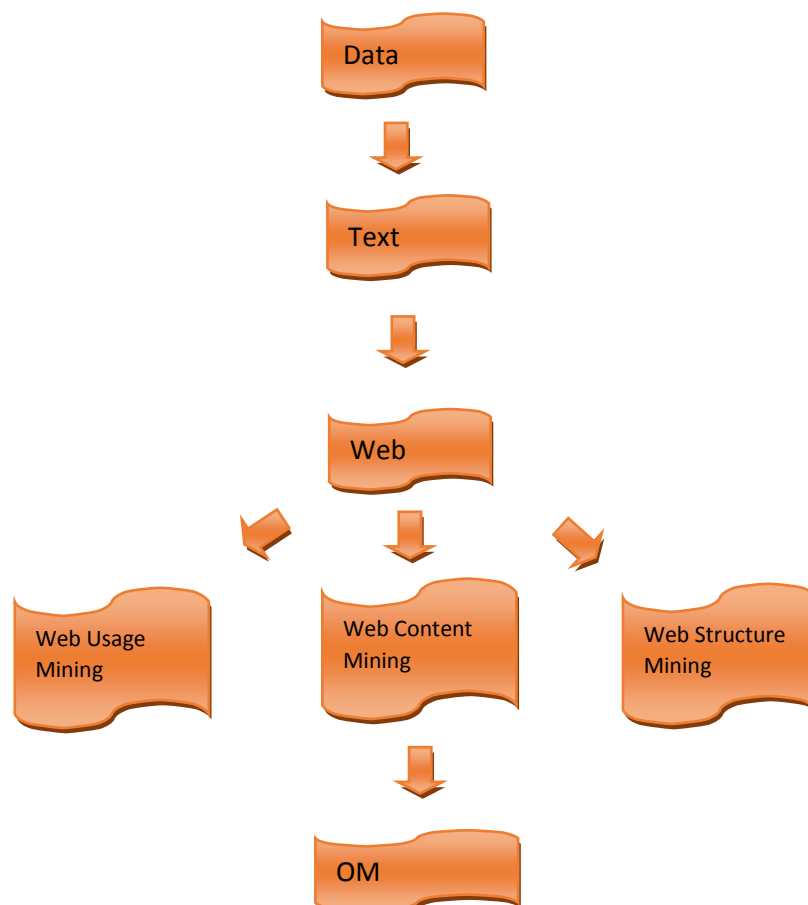


Figure 2: Hierarchy of Data Mining

2.5. Opinion Mining or Sentiment Analysis

Everyday many users purchase products and services through online services in the web. For most of the online buyers, it has been a habit to comment, write reviews on the purchased product or service. Therefore, most of the buyers tend to read the comments and reviews about the prospective buying item and make decisions based on them. However, the problem is, because of the large number of comments and reviews it has been difficult to go through all and clearly identify them. Therefore, it needed a mechanism to go through large no of reviews and extract the useful information in a short period of time. Opinion mining comes in to play in this scenario and important aspect in opinion mining research has been sentiment analysis (Alexandra Cernian *et al.*, 2015). Opinion mining can be classified in to two broad faces – Opinion Structure and Opinion Mining Tools and Techniques (Rajeev Aurora and Srinath Sirinivasa, 2014). Opinion mining is commonly used in academia and sentiment analysis is in industry (Chinsha T C and Shibliy Joseph, 2015).

It is important to get feedback from clients and prospective clients to decision-making process, get suggestions for product improvement, adapting market strategies.

An opinion can be described as the emotional words or some attitude about one topic. Opinion can be one of the three categories below.

- Positive
- Negative
- Neutral
- Constructive - Suggestions to improve or make the product or service better. Does not imply that the opinion holder is negatively inclined about the entity (Rajeev Aurora and Srinath Sirinivasa, 2014).

G. Angulakshmi et al. (2014) categorized opinions in to two main categories as direct opinions and comparisons. Rajeev Arora et al. (2014) talks about indirect opinions apart from the above mentioned two categories.

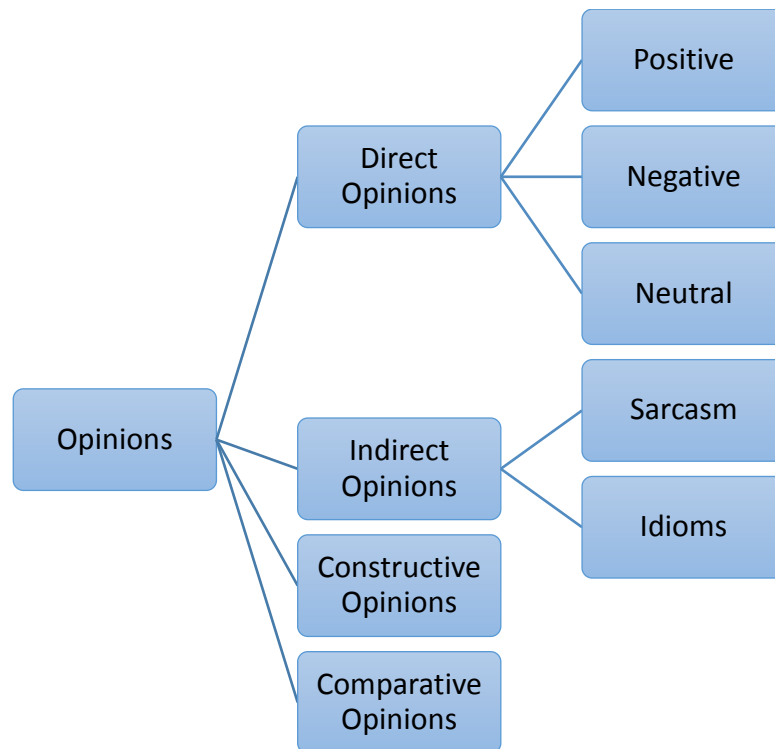


Figure 3: Types of Opinions

Opinion mining is classification of user's expressed opinion into positive or negative polarity (Vijay B. Raut *et al.*, 2014). As G. Angulakshmi *et al.* (2014) opinion mining is a technique, which used to detect and extract subjective information in text documents. OM was not concerned in traditional text classification (Juling Ding *et al.*, 2009). Aim of OM is to automatically extract opinions expressed in user-generated content (A. Jeyapriya, 2015). Opinion mining is a hard problem because it is difficult to automate interpreting the meaning of a sentence. However, the results can be very useful with the increasing usage of web.

Opinion mining is a very popular research area in natural language processing and topic in text mining. As Chinsha T C and Shibliy Joseph (2015) opinion mining is a sub problem of traditional text classification problem. Initial research in text mining focused on factual information from documents, which are typically objectives sentences. But now the focus is shifting towards opinion mining which is identifying opinions, which are subjective in nature (Rajeev Aurora and Srinath Sirinivasa, 2014), and there are many applications developed based on it. The original opinion mining area could only classify a document/text to be "subjective" or "objective" which does not give any practical significance. Then the sentiment classification about "positive", "negative", "neutral" which is called attitude type mining is developed. Then the researches have been carried out to identify the strength of the attitude (Juling Ding *et al.*, 2009).

Opinion mining can be done in three main levels, document level, sentence level and phrase level (M. Lovelin Ponn Felciah and R. Anbuselvi, 2015), (Chinsha T C and Shibily Joseph, 2015), (Amani k Samha, 2016), (A. Jeyapriya et al., 2015). It has been difficult to identify the exact meaning of reviews through document and sentence level OM. Therefore, the current OM research is moving towards phrase level OM (Alexandra Cernian *et al.*, 2015) which is also known as aspect based OM. Aspect based OM is identifying most important aspects of an item and identify the positivity and negativity using those aspects automatically. As M. Lovelin Ponn Felciah and R. Anbuselvi, (2015) feature based sentiment classification has become a predominant area in opinion mining and it considers certain subjects feature opinions.

Opinion mining can be classified in to following categories based on the method used for mining.

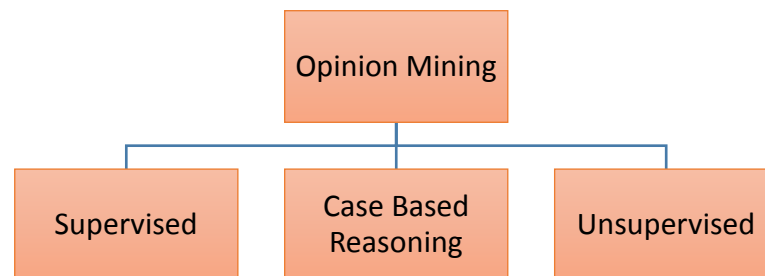


Figure 4: OM Categories Based on the Method Used in Mining

Knowledge gained from opinion mining helps people to identify better products and services in the market and make effective decisions when purchasing. For the businesses, it helps to identify products opinions, brand view, level of reputation etc. (A. Jeyapriya, 2015).

2.6. Opinion Mining Approaches

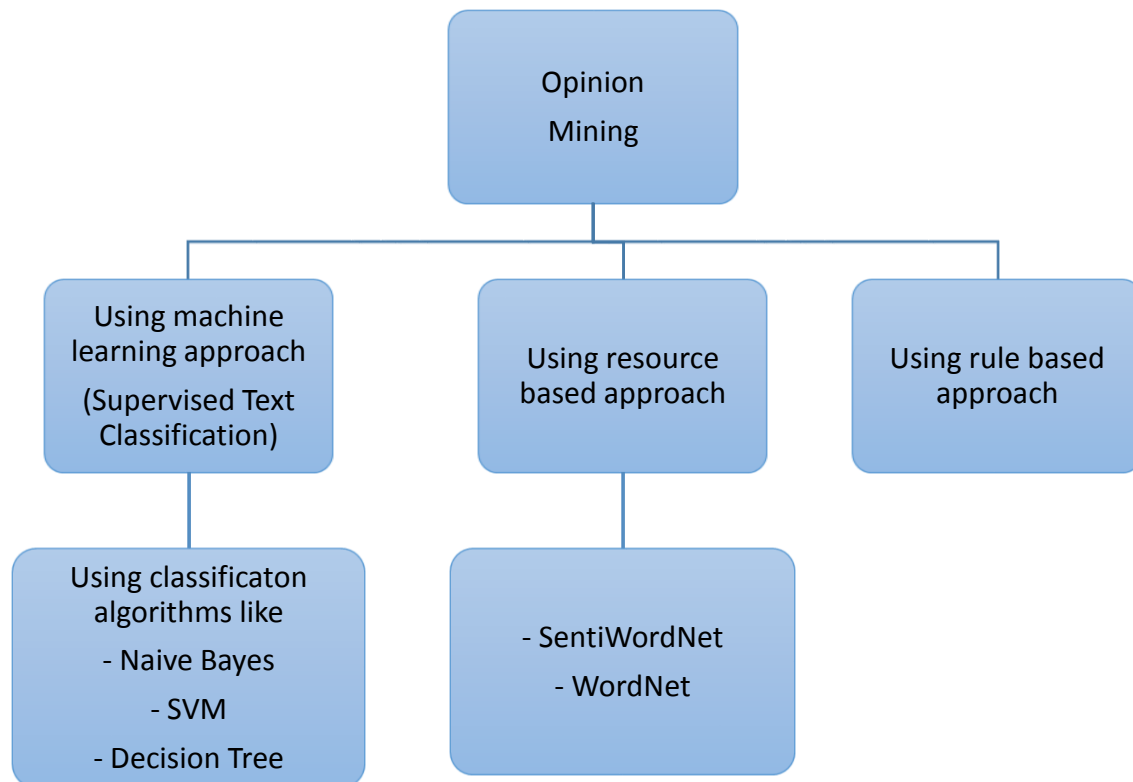


Figure 5: OM Approaches

Product review classification is the main objective of Opinion Mining and various techniques have been used to address it. As Alaa Hamouda et al. (2015) and Rajeve Arora et al. (2014) opinion mining approaches can be classified in to two main categories, as approaches based on lexical resources and natural language processing and approaches using machine learning algorithms. G. Angulakshmi *et al.* (2014) divide opinion mining techniques to three main categories, Supervised Machine Learning, Unsupervised Learning and Case based reasoning.

Following are the summarization of main approaches.

Approach	Supervised machine learning approach
Description	<p>Algorithms are trained on sample labeled review text to build a classifier model. (Vijay B. Raut <i>et al.</i>, 2014)</p> <p>Best performance has been obtained using support vector machines in combination with unigrams. (Alaa Hamouda et al. 2015)</p> <p>Most frequently used popular data mining technique (G. Angulakshmi et al., 2014).</p>

Advantages	Disadvantages
1. Not language specific	1. Need an annotated training data set (Rajeev Aurora and Srinath Sirinivasa, 2014)

Approach	Case based approach
Description	Emerging artificial intelligence supervised technique. (G. Angulakshmi et al. ,2014)
Advantages	Disadvantages
	1. Need a knowledge repository.

Approach	Unsupervised Machine learning approach
Description	Has no explicit targeted output associated with input. Learned by observation. Clustering is the technique used. (G. Angulakshmi et al., 2014).
Advantages	Disadvantages
	1. Need a larger data set

Approach	Statistical algorithms with a lexicon or syntactic based approach with a lexicon
Description	Lexicon is a language-specific dictionary of words with a predefined polarity.
Advantages	Disadvantages
1. No need of any training data.(Chinsha T C and Shibily Joseph, 2015)	As Rajeev Aurora and Srinath Sirinivasa, 1. A lexicon is language specific 2. Word polarity is domain specific. For example, quiet refers to a positive sentiment in a car while it is a negative sentiment for a phone

	3. Polarity is also context specific within a domain. For example, long is a positive sentiment for battery life of a laptop but it refers to a negative sentiment for startup time of the laptop.
--	--

2.7. Opinion Mining Levels

Level	Document level
Description	<p>Find an overall sentiment score at the document level.</p> <p>Mainly formulated as classification problems where input document is classified in to few predefined categories (A. Jeyapriya et al., 2015).</p> <p>Here single review about a topic is considered (G. Angulakshmi et al., 2014).</p>
Advantages	Disadvantages
	<ol style="list-style-type: none"> 1. Assume a document contains an opinion for only one entity and aspect (Rajeev Aurora and Srinath Sirinivasa, 2014). 2. Information are not sufficient for valuable decision-making process (A. Jeyapriya et al., 2015).

Level	Sentence level
Description	<p>Each sentence is classified as subjective or objective and then the polarity of each of the subjective sentences are inferred (M. Lovelin Ponn Felciah and R. Anbuselvi, 2015), (G. Angulakshmi et al., 2014).</p> <p>Sentences are retrieved and ranked based on some criteria (A. Jeyapriya et al., 2015).</p>
Advantages	Disadvantages
	<ol style="list-style-type: none"> 1. Assumes each sentence contains an opinion for one entity and aspect, aspect (Rajeev Aurora and Srinath Sirinivasa, 2014)

	2. Information are not sufficient for valuable decision-making process (A. Jeyapriya et al., 2015).
--	---

Level	Aspect level
Description	<p>Discovers entities, aspects and sentiments of each of them. Aspects can be explicit or implicit. (Aurora and Srinath Sirinivasa, 2014).</p> <p>Aspect based opinion mining extract most important aspects of an item and then predicts the orientation of each aspect from the item reviews. (A. Jeyapriya, 2015)</p> <p>Core tasks in aspect based opinion mining is aspect identification, aspect based opinion word identification and its orientation detection (Chinsha T C and Shibliy Joseph, 2015), (Alexandra Cernian <i>et al.</i>, 2015), (Lizhen Liu et al., 2012).</p>
Advantages	Disadvantages
<ol style="list-style-type: none"> 1. More information for decision making 2. Assumes a document contains opinion on several entities and their aspects. (Rajeev Aurora and Srinath Sirinivasa, 2014) 	<ol style="list-style-type: none"> 1. Need to handle negations G. Angulakshmi et al. (2014).

2.8. Opinion Mining Process

Opinion mining process is described by several researches and according to Aurora and Srinath Sirinivasa (2014), M. Lovelin Ponn Felciah and R. Anbuselvi (2015), (Monalisa Ghosh et al., 2013) it can be described using below steps.

Aspect identification

Aspects are the important features rated by the reviewers (Chinsha T C and Shibliy Joseph, 2015). Aspects/features can be categorized in to two main categories, implicit features and explicit features (Lizhen Liu et al., 2012). Normally explicit aspects are noun or noun phrases around the base entity (Amani K Samha, 2016), (Rajeev Arora et al., 2014). Rajeev Arora (2014) and (Lizhen Liu et al., 2012) emphasize that frequently used phrases as key aspects of the entity while infrequent

are less important aspects. In feature level opinion mining the most important task is to identify the features (Lizhen Liu et al., 2012). As described by Chinsha T C and Shibliy Joseph, (2015), M. Hu and B. Liu (2004), A. Jeyapriya et al. (2015), Rajeev Arora (2014) used association rule mining with pruning strategies to find the candidate features using below steps,

- Perform POS (Parts of Speech) parsing – a dedicated parts of speech tagging software has been used in the application proposed by Alexandra Cernian *et al.* (2015)
- Filter noun and noun phrases
- Fed the nouns to association rule mining algorithm

Apart from the above method J.S. Kessler and N. Nicolov (2009) introduced machine learning classifier (SVM) to find related opinion expression and target aspect and G. Qiu, B. Liu, J. Bu and C. Chen used double propagation method. Lizhen Liu et al. (2012) introduced a totally different process to extract aspects or features by taking adjectives as the opinion words and the use those adjectives to extract corresponding features. If there are no feature found it would assume there are one implicit feature for it.

Sentiment word identification

After finding aspects, the next task is to find sentiment words. Opinion words are the words, which express opinion towards aspects. Opinion words and phrases are normally adjectives, which can be extracted using POS tagging tools (G. Angulakshmi et al., 2014), (Rajeev Arora 2014). However, Chinsha T C and Shibliy Joseph, (2015), proposed a more improved sentiment word identification using adjectives, verbs, adverb adjective and adverb verb combinations. Then a syntactic based approach has been used with the help of Stanford Dependency parser. G. Angulakshmi et al. (2014) also talks about adjective and adverb combination when extracting features.

Sentiment orientation

Aurora and Srinath Srinivasan (2014) talks about two methods, PMIIR score based method and semi supervised learning method that uses a seed list to find sentiment orientation. A. Jeyapriya et al. (2015) used Naïve Bayesian algorithm using supervised term counting based approach to identify the number of positive and negative opinions of each extracted aspect. Rajeev Arora (2014) has paid attention on the connective words like AND – use to connect either positive or negative sentiments or disjunctions – BUT, OR, EITHER, OR – use to join two dissimilar opinions when finding sentiment orientation.

Chinsha T C and Shibliy Joseph, (2015) proposed an opinions mining process and they extended the process by adding few important like review collection, preprocessing – stop word removal, subjective sentence identification, aspect level score calculation etc. which are described below.

- Review collection -

Reviews are input for the opinion mining process. Web crawlers can be used as information retrieval technique (G. Angulakshmi et al., 2014).

- Preprocessing -

Preprocessing is clean the reviews which improve the accuracy of opinion mining process which avoid the unnecessary processing overhead. Alaa Hamouda and Mohamed Rohaim(2015) proposed a tokenization process, which splits the text in to very simple tokens as a preprocessing step other than the preprocessing steps like POS tagging etc. G. Angulakshmi et al. (2014) emphasized case normalization as a preprocessing step in his literature.

- Subjective sentence identification -

Identify the opinionated sentence and remove objective sentence, which does not contain an opinion.

- Aspect level score calculation

The polarity score of an aspect in a sentence is calculated by aggregating opinion words scores in that sentence. Here priority scores are assigned using SentiWordNet, which is a dictionary of sentiment words.

Amani k Samha (2016) further improved the process by adding below preprocessing steps to the process.

- Group aspect synonyms

Various people can use various phrases or words to refer the same aspect.

E.g.: - display, LCD, screen,

- Lemmatization process from the beginning of preprocessing.

A. Jeyapriya (2015) also proposed stemming as a preprocessing step, which is forming root word of a word. He has used Porter Stemmer algorithm to form root word for given input.

Juling Ding et al. (2009) proposed a flexible opinion mining model based on opinion tree, which is a totally different process than above. However, it has become only a theoretical model, which does not have a practical significance.

2.9. Opinion Summarization

Opinion summarization is the process of generating effective summaries of opinions, which is helpful for the users to get quick understanding by looking at them. As Vijay B. Raut *et al.* (2015), Opinion summarization is the process of finding most important aspects and representation of review information in short and summarized form. As M. Lovelin Ponn Felciah and R. Anbuselvi, (2015) this is different from traditional summarization. Summaries can be in sentence level or aspect level. Opinion summarization is a major part in OM process (G. Angulakshmi et al., 2014).

G. Angulakshmi et al. (2014) describes two main approaches in opinion summarization, which are Feature based summarization, and term frequency based summarization.

2.10. Opinion mining tools

OM tools allow business to understand new product opinion, product sentiments, brand view and reputation management. Opinion mining tools and techniques can be explained as methods to extract opinion structure and aggregation of sentiments (Rajeev Arora, 2014). These tools help users to perceive product opinions or sentiments in global scale (A. Jeyapriya et al., 2015). Review Seer tool, Web Fountain, Red Opal and Opinion observer are some of the developed OM tools (G. Angulakshmi et al., 2014).

2.11. Study of Past Research

Research	Vijay B. Raut, D.D. Londhe(2014)
Features	<ul style="list-style-type: none"> • Presents a machine learning and SentiWordNet based method for OM and sentence level score based method for opinion summarization • Proposed system – <ul style="list-style-type: none"> ○ Review text retrieval – from review web sites like www.tripadvisor.com by crawling techniques.

	<ul style="list-style-type: none"> ○ Classification - as positive or negative using machine learning classifiers – Naïve Bayes, Support Vector Machine, Decision Tree etc. and SentiWordNet based algorithm after preprocessing ○ Summarization - using sentence extraction method. Most informative sentences and most relevant to the context are selected from document to represent the summary. Review sentences are scored and sorted based on the score calculated. ○ Classification module is tested using WEKA library on parameters like Precision, Recall and F-measure.
Pros	<ul style="list-style-type: none"> • Implemented system has been tested against several algorithms for accuracy, which is helpful identifying the best algorithm.
Cons	<ul style="list-style-type: none"> • Has not focused on preprocessing which is a very important process in OM before sending the text in to “Classification”
Improvements	<ul style="list-style-type: none"> • Conduct preprocessing activities before OM.

Proposed System

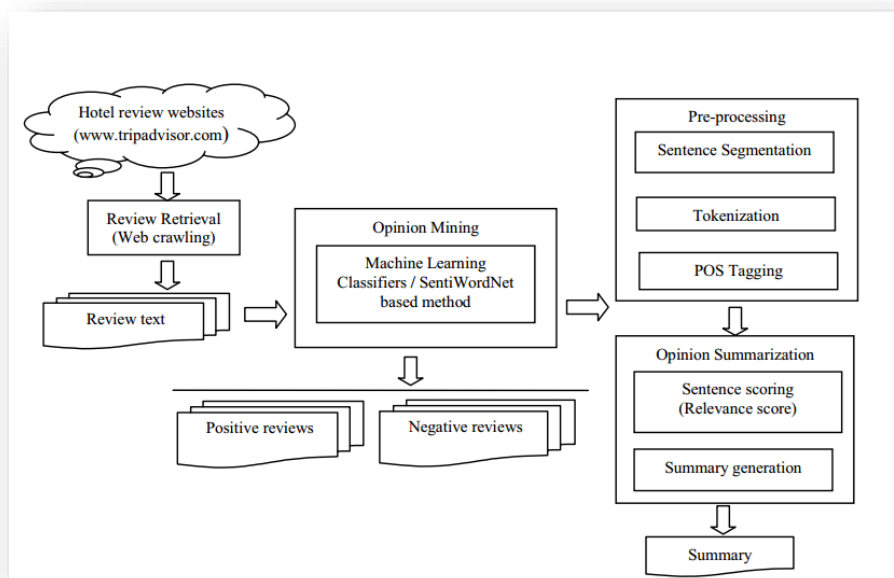


Figure 6: Proposed System Architecture

Research	A Jeyapriya, C.S.Kanimozhi Selvi(2015)
Features	<ul style="list-style-type: none"> • Phrase level OM is considered • Proposed process <ul style="list-style-type: none"> • Review collection - from the sites like www.amazon.com, www.epinions.com and www.cnet.com. • Stop Word Removal - remove stop words like is, are, was etc. • Stemming - form root words of a word using Porter stemmer algorithm. • POS Tagging - label each word in a sentence with its appropriate part of speech – verb, noun, adjective, adverb, pronoun, preposition, conjunction, and interjection (linguistic category that is defined by its syntactic or morphological behavior) using Stanford tagger. • Aspect Extraction - mine frequent item sets • Sentence and Aspect Orientation - <ul style="list-style-type: none"> ○ Determine the number of positive and negative opinion sentence in reviews using opinion words. ○ Identify the number of positive and negative opinions of each extracted aspect. ○ Implement sentence and aspect orientations using Naïve Bayesian algorithm using supervised term counting based approach. ○ Calculate the probabilities of the positive and negative count using Naïve Bayesian classifier. • Performance evaluation - using Precision, recall and F-measure.
Pros	<ul style="list-style-type: none"> • Focused on important factors in preprocessing like, POS tagging, aspect extraction etc. • Applied a rule for negation handling which reverse the positive or negative count. • 80% accuracy using frequent item set mining and 92% accuracy using sentiment orientation

	<ul style="list-style-type: none"> Has used precision, recall and F measure to system evaluation.
Cons	<ul style="list-style-type: none"> The sentence or opinion is marked as below, <ul style="list-style-type: none"> Positive - if the probability of positive count is greater than the negative counts Negative - if the probability of negative count is greater than the positive counts Neutral - if the probability of positive count minus probability of negative count is zero <p>Calculation method will be suitable for simple sentences but for the complex sentences it will not be sufficient. The value for positive opinion may not be equal to negative value.</p>
Improvements	<ul style="list-style-type: none"> Relative importance of the extracted aspect should be considered. (Score based on sense number).

Proposed System

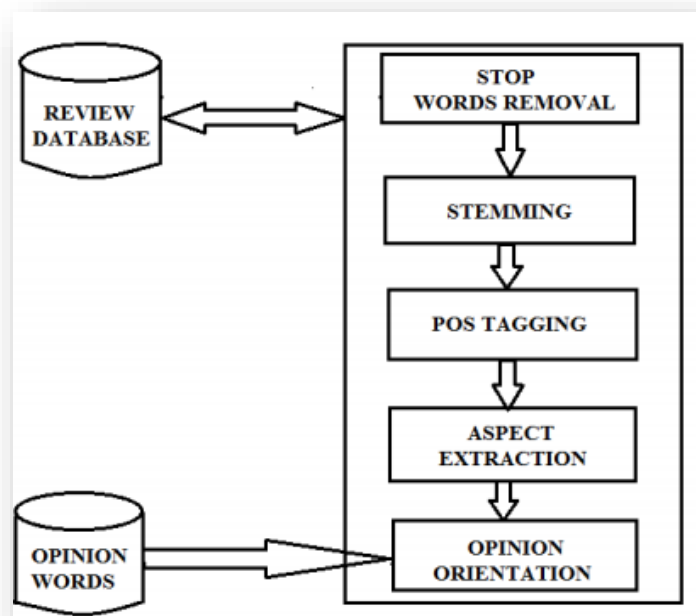


Figure 7: Working of Proposed System Architecture

Research	Juling Ding, Zhongjian Le, Ping Zhou, Gensheng Wang, Wei Shu (2009)
Features	<ul style="list-style-type: none"> Proposed a new kind of tree which is called opinion tree Root node is the subject of the public opinions. The first layer node is opinion of coarse granularity – positive, negative, neutral. The second layer node is medium size opinion – high, medium, low. The leaf node is fine-grained opinion.
Pros	<ul style="list-style-type: none"> Talks about the attitude force which is high, medium, low
Cons	<ul style="list-style-type: none"> Only a theoretical approach and does not provide any practical process to implement the model.

Proposed Opinion Tree

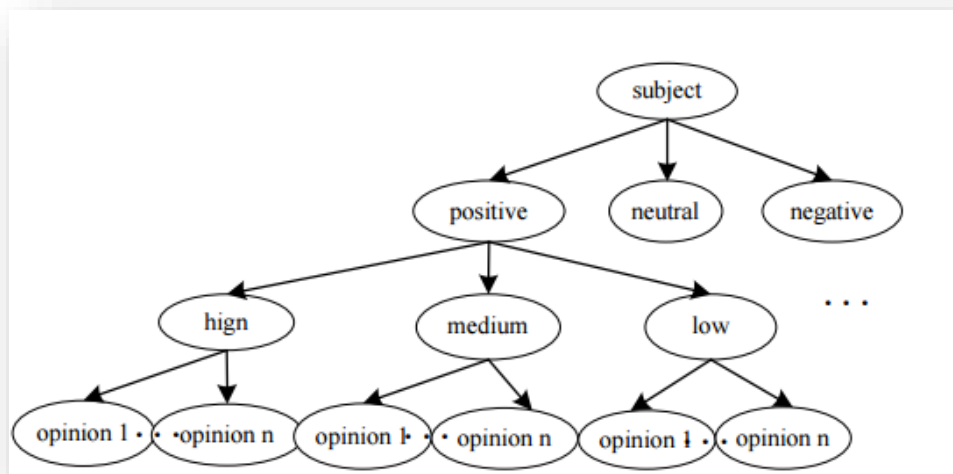


Figure 8: Opinion Tree

Research	Rajeev Arora, Srinath Srinivasa(2014)
Features	<ul style="list-style-type: none"> Frequently used phrases are taken as key aspects of the entity while infrequent are less important aspects Used Apriori algorithm to extract aspects Paid attention on the connective words like AND – use to connect either positive or negative sentiments or disjunctions – BUT, OR, EITHER, OR – use to join two dissimilar opinions.

	<ul style="list-style-type: none"> Talked about cross-domain classification, sarcasm detection and spam detection.
Pros	<ul style="list-style-type: none"> Makes our attention to several areas that need to be considered during OM
Cons	<ul style="list-style-type: none"> Not elaborated details on them or has not mentioned about the methods or process to implement those points

Research	Chinsha T C , Shibily Joseph(2015)
Features	<ul style="list-style-type: none"> Focus on aspect level opinion mining Proposed a new syntactic based approach using syntactic dependency, aggregate score of opinion words, SentiWordNet and aspect table Proposed opinion mining process, <ul style="list-style-type: none"> Reviews collection – reviews are downloaded from a review downloader and stored in a database Aspect extraction - done by identifying the noun and noun phrases after POS tagging the sentences Preprocessing - includes some preprocessing steps like stop word removal, subjective sentence identification Subjective sentence identification Opinion words identification - done via processing the adjectives verbs, adverb adjective combinations and adverb verb combinations Aspect level score calculation - aspect level scores are calculated aggregating opinion words scores in the sentence. Positive and negatives scores for a synset in SentiWordNet is used as scores of the opinion words Calculation of aspects score in all reviews – total score of an aspect from all reviews is find out by aggregating the sentence wise score of that aspect Negation handling has been done by reversing the scores if a negative relation found

	<ul style="list-style-type: none"> Performance evaluation has done with Precision and Recall. Total accuracy has been 78.04%.
Pros	<ul style="list-style-type: none"> Has covered most of the required elements in OM application like preprocessing, aspect identification, negation handling etc. Accuracy rate is above 75%
Cons	<ul style="list-style-type: none"> If an opinion word is present more than one time in SentiWordNet, the average of highest positive or negative score has been taken
Improvements	<ul style="list-style-type: none"> Average may not be the suitable method since the weight of the opinion words can be different. Therefore, in this work a mechanism will be implemented to calculate weighted average score and used it as the opinion word score.

Research	Alexandra Cernian, Valentin Sgarciu, Bogdan Martin(2015)
Features	<ul style="list-style-type: none"> Presents a sentiment analysis using SentiWordNet Proposed opinion mining process, <ul style="list-style-type: none"> POS tagged using Stanford POS tagger Split review phrases in to sentences and sentences in to words Calculate sentence score using below formula $\text{sentence_score} = \frac{\sum_{i=0}^n \text{item_score}(i)}{\text{no_words}} * \text{size_index}$ <p>Size index is assigned according to the length of the sentence.</p> $\text{text_score} = \frac{\sum_{i=0}^n \text{sentence_score}(i)}{\text{no_of_sentences}}$
Pros	<ul style="list-style-type: none"> A rating system is proposed based on the scores A length index has been introduced according to the length of the sentence.
Cons	<ul style="list-style-type: none"> Item score calculation mechanism is not properly described

Research	Lizhen Liu, ZhixinLv, Hanshi Wang(2012)
Features	<ul style="list-style-type: none"> • Features are identified via adjectives and clustered the same features. • Features have been identified in the opposite way of all other researches. <ul style="list-style-type: none"> ○ first identifies the adjectives as opinion words ○ then uses them to extract features - noun / noun phrase, verb / verb phrase. ○ if there is no feature found for the adjective marks the adjective for implicit features.
Pros	<ul style="list-style-type: none"> • Consider about both explicit features and implicit features.
Cons	<ul style="list-style-type: none"> • Mainly focuses on the Chinese comments

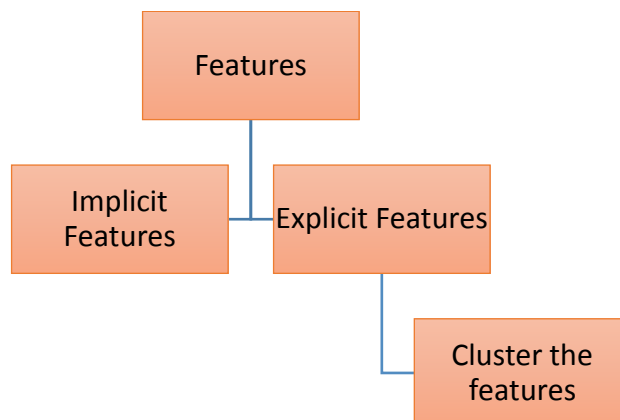


Figure 9: Feature Categories

Research	Alaa Hamouda, Mohamed Rohaim(2015)
Features	<ul style="list-style-type: none"> • Provides an improvement for previously used approaches using SentiWordNet
Pros	<ul style="list-style-type: none"> • Calculate positive and negative scores by getting the average for its entries according to POS.
Cons	-
Improvements	<ul style="list-style-type: none"> • Implement weighted average.

Research	Monalisa Ghosh, AnimeshKar(2013)
Features	<ul style="list-style-type: none"> Proposed a technique for sentence level sentiment classification using rule based method. Proposed process <ul style="list-style-type: none"> Preprocessing - remove stop words, spelling correction Review sentence tagging - POS tagging the sentence using Stanford POS tagger Feature Extraction- retrieve frequent nouns, extract adjectives Phrase extraction - extract noun phrase by calculating the position between frequent feature and opinion word Sentence classification using SentiWordNet - get positive or negative score of first lemma for aspect score and calculate the sentence score by averaging the total aspect score. Calculate the review score by averaging the sentence score.
Pros	<ul style="list-style-type: none"> Described a good process and better score calculation methods
Cons	-
Improvements	<ul style="list-style-type: none"> Has only considered the first lemma for aspect score calculation. But in the proposed system an improvement will be done by taking all the terms not only the first lemma

Research	Amani K Samha (2016)
Features	<ul style="list-style-type: none"> Consist of two major steps <ul style="list-style-type: none"> Pre processing <ul style="list-style-type: none"> Remove symbols like {, [, :], :(..., s Stanford Lemmatization Split to sentences Run the Stanford Dependency Relations to find the syntactic parsers Main processing

	<ul style="list-style-type: none"> ▪ Aspect extraction – assume nouns are aspect candidates ▪ Opinion extraction – assume adjectives are opinion candidates and use syntactic rules to find aspects and opinions ○ Used precision, recall and F-measure to evaluate the efficiency
Pros	<ul style="list-style-type: none"> • Achieved better accuracy than the existing aspect based opinion mining models • Evaluation methods has been used for a proper understanding on the results
Cons	-

When considering all the researches done so far, following factors are taken into consideration when implementing an OM application.

- Document level and sentence level OM are not enough to gather true insight of reviews. Therefore, it is essential to perform an aspect based or phrase level opinion mining to find valuable knowledge from the reviews (Chinsha T C and Shibliy Joseph, 2015). Therefore, aspect based opinion mining in sentence level is preferred in here.
- As Rajeev Arora Srinath and SrinathSirinivasa (2014) frequent item set mining will be done using Apriori algorithm and adjective opinion words are identified using POS tagging tool.
- Chinsha T C and Shibliy Joseph (2015) has calculated the average of highest positive or negative score if an opinion word is present more than one time in SentiWordNet. However, average does not care about the relative importance of the word and therefore in this work, weighed average calculation method is proposed. Then it will not take same positive or negative value for all sysnet consideration when calculating scores of the sentence or opinion.
- Negation handling which is the process of identifying negation words like “not”, ‘don’t’, ‘never’ and consider them when calculating the OM scores is needed for an efficient application. If not negative, reviews can be categorized and positive sentence and it will end up decreasing the accuracy rate and efficiency of the application. The negation handling

mechanism proposed in Chinsha T C and Shibliy Joseph (2015) will be used when implementing the application.

- The methodology proposed by Chinsha T C and Shibliy Joseph(2015) will be used in opinion words identification task as it has improved the traditional opinion word identification methodology using adjectives by verbs, adverb adjective and adverb verb.
- Preprocessing is a mandatory process in OM. It will clean the data set and make the review text for the desired format for OM. Following preprocessing activities are important as Chinsha T C and Shibliy Joseph(2015), Amani K Samha(2016) and they will be done in this work. Preprocessing tasks that will be implemented in this project.
 - Remove unwanted characters
 - Stop word removal
 - Stemming
 - POS tagging
 - Remove duplicate author comments
- In OM, supervising learning machine algorithms needed annotated training before performing the initial run. However, in SentiWordNet does not. SentiWordNet is practically developed for English language and other few languages like Chinese, German. It is the better approach of OM without any prior data set. In addition, recent studies have shown the Natural Language Processes NLP techniques based on dependency relations enhance the accuracy and performance of unstructured prediction problems (Amani K Samha, 2016).
- As Juling Ding et al. (2009), it needed to identify the attitude force or strength of the opinion whether it is high, medium or low for a better result. This will be partially handled by SentiWordNet as it assigns different scores of positivity or negativity based on the word.
 - E.g.: - positivity score of “best” is greater than the positivity score of “good”
- Rajeev Arora Srinath and Srinath Sirinivasa, (2014) researcher emphasize on the facts like, constructive opinions (improvements) and sarcasm detection. Improvements should not be taken as negative opinions. In addition, it would be great if we can identify sarcasm opinions. However, the problem is it is complex to identify them.

Therefore, those are not taken as the properties of the prospective system that is being developed. They can be taken as the requirements for next phase of research and development.

Also in this paper, it makes our attention to credibility of an opinion. Trust and credibility of an opinion depends on the person who is giving the opinion and what is his motivation. For example, a competitor of a product might put negative reviews in Facebook post purposely to create a negative impact of future buyers. In addition, businesses can post lot of positive comments about their product and services, (opinion spamming). However, these kinds of situations are difficult to identify and handle in automatic processes since author behaviors patterns are not mind in these kinds of OM applications. Therefore, such scenarios will not be in the scope of this project.

- Apart from the score calculation method for opinion words like average score, we can introduce a weighted average score calculation method in our prospective application.
- As Alexandra Cernian, Valentin Sgarciu and Bogdan Martin (2015) rating system should be developed based on the sentences scores calculated while processing the opinion words.
- As Lizhen Liu, ZhixinLv and Hanshi Wang(2012) frequent item sets of nouns in a comment chooses as product features and infrequent item sets are ignored in the OM process. However, those ignored items are also product features and may give valuable information about the product. This is a valid point and should be considering in future work.
- Monalisa Ghosh and AnimeshKar (2013) described a good flow and an improvement for score calculation methods. Therefore, these will be considered in the proposed system.
- In the proposed system, only explicit aspects are considered as the process could not identify implicit aspects as mentioned by (Chinsha T C and Shibily Joseph, 2015).
- Previous research commonly performed lemma tagging at the end, but we did the lemma at the beginning of the process, aiming to group similar words to find frequent aspects and opinions treat them as single item (Amani K Samha, 2016).

- To evaluate the accuracy precision, recall and f measure will be used in as (Amani K Samha, 2016).

2.12. Summary

In this chapter, firstly an analysis has been done on the key words of the study like Social media, data mining, opinion mining etc. Then the focus has been put on the challenges in this research area. After that a detail study has been done on the past researches by going through more than 10 good research papers and every research has been criticized with the pros, cons and possible improvements. Finally, a critical evaluation is presented considering all the research papers by selecting the key points included in the papers.

3. Research Methodology

Introduction	
Research Methodology	
Development Methodology	
Project Management Methodology	
Summary	

3.1. Introduction

This chapter describes the methodologies used in the project. Research methodology, development methodology and project management methodology are described along with the reasons to select them.

3.2. Research Methodology

As Dr. S. M. Aqil Burney (2008), research methodology can be categorized in to two main approaches, inductive and deductive. In this project, deductive research approach that works from more general to more specific and also a top down approach has been used. Initially extensive study was carried out on all the theoretical approaches to implementing opinion mining applications and then most suitable ones are selected for further study. Finally, the application built on top of the theories studied deeply and extensive testing has been carried out to confirm them.

As Saul McLeod (2008), research methods also can be categorized as qualitative and quantitative. In this project, both qualitative and quantitative methods have been used to gather data. Qualitative methods like questioners, interviews are carried out find out what other people think about the project in evaluation phase. In the testing phase, various measures like Precise, Recall, F Measure etc. have been calculated to find the accuracy of the project output, which is a quantitative method.

3.3. Development Methodology

With the advancement of software development industry, various development methodologies have been created to improve the process. Following is a summarization of famous development methodologies today,

	Pros	Cons
Waterfall Model	Easy to understand Easy testing and analysis	Rigid model Does not allow editing in the middle
Agile Methodology	Adoptive approach Allow changes	Low documentations
Rapid Application Development Methodology	Provide quick results	Depend on the team performance

Spiral Model	Low risk Suitable for large scale projects	Costly
Extreme Programming Methodology	Customer involvement	Requires frequent meetings
Feature Driven Development	Easy development	Not suitable for smaller projects and single developer

By going through all the development methodologies and their advantages and disadvantages Agile is selected in this project, which is an “iterative” and “incremental” process. As a research project methodology should allow changes. Also as findings, progress development methodology should support for iterative and incremental process. Therefore, without building the whole project at once, smaller modules like UI module, preprocessing module – review collection, stop word removal, stemming etc. are developed in iterations.

e.g.: -

- Requirement analysis for stop word removal
- Design stop word removal function
- Implement it
- Test it
- Fix issues found in stop word removal function
- Re test it
- Start the other function

3.4. Project Management Methodology

As Mike Wooldridge Good project management cannot guarantee success, but poor management on significant projects always lead to failure. Project management is essential in software development projects because of the high risk. Time, cost and quality are the triple constraints in software projects.

Following are some of the widely used project management methodologies today.

	Description
PRINCE2	Widely used in private sector Broad collection of good practices
Kanban	Project work is displayed on a board Visual display on what is coming up next

Six Sigma	Data driven product and process improvement methodology Improve process by eliminating defects
DMAIC	Part of Six Sigma methodology Often used as a standalone method

In this project, PRINCE2 is used as the project management methodology and following summarize the time and risk management process used.

3.4.1. Time Management

Main phases of the project like literature review, requirement analysis, designing, testing etc. are identified at the beginning of the project and time is allocated for each of them identifying the sub tasks as well. Each of the phase consist of a time line and a deliverable. Refer appendix for the Gantt Chart, which gives detail a detail explanation on time allocation of the project.

3.4.2. Risk Management

Every software inherits a risk and since this is a research project high risk is associated with the project.

Following are some of the risks identified and details of mitigation plan.

Risk	Lack of latest academic knowledge
Risk Level	High
Description	Opinion mining is a novel research area for me and the prior knowledge about the area was less. Also it a fast growing research area and everyday new knowledge introduce to the world.
Mitigation Plan	Conduct extensive search on latest research papers in literature review phase.

Risk	Find free java supporting libraries
Risk Level	High
Description	In this opinion mining application lot of preprocessing activities are performed. Due to time constraints, it is hard to implement those functions in this project. Therefore, java libraries like POS tagging libraries, data pruning libraries etc are needed to implement the system.
Mitigation Plan	Conduct extensive search in internet.

	<p>Talk with colleagues to share knowledge.</p> <p>Compile free and open source existing java classes and used them in the project.</p>
--	---

Risk	Find test data
Risk Level	High
Description	To test the implementation a large data set is needed. It is time consuming to manually retrieve the review comments from social media sites.
Mitigation Plan	<p>Conduct extensive search for data sets.</p> <p>Prepare a process to clean the available data sets.</p>

Risk	Loss of data
Risk Level	High
Description	Since all the project related data in laptop, there can be a chance to loss it due to hardware failures.
Mitigation Plan	Take a backup at every week and keep it in a pen drive.

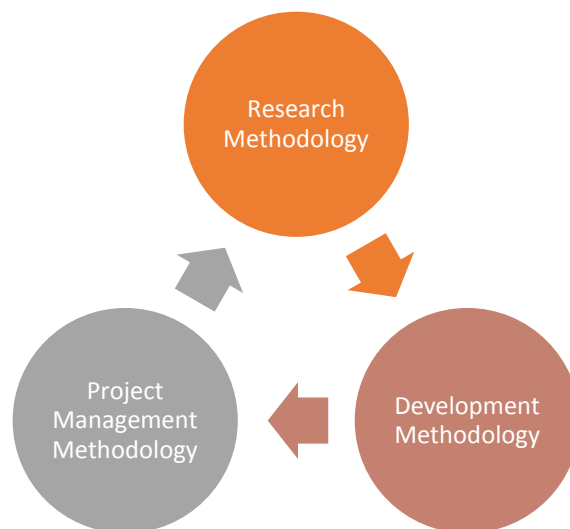


Figure 10: Research Methodology

3.5. Summary

The chapter has described the research methodology used in the project and reasons to use them. Every section describes the available methodologies in the industry, methodology used in this research project and detail description for the reasons to use in the current project.

4. Requirement Specification

Introduction	
Rich Picture	
Context Diagram	
Stakeholder Analysis	
Requirement Analysis	
Use Case Models	
Use Case Diagram	
Use Case Description	
Summary	

4.1. Introduction

This chapter describes the proposed system's functional, nonfunctional requirements, use cases etc. and the complete description how the system is being performed.

4.2. Rich Picture

The Rich Picture (RP) is a flexible graphical technique, which may be used as part of the Checkland Soft Systems Methodology (SSM) (Pat Horan, 2000). Following is the rich picture of the problem, which Opinion Mining application is going to resolve.

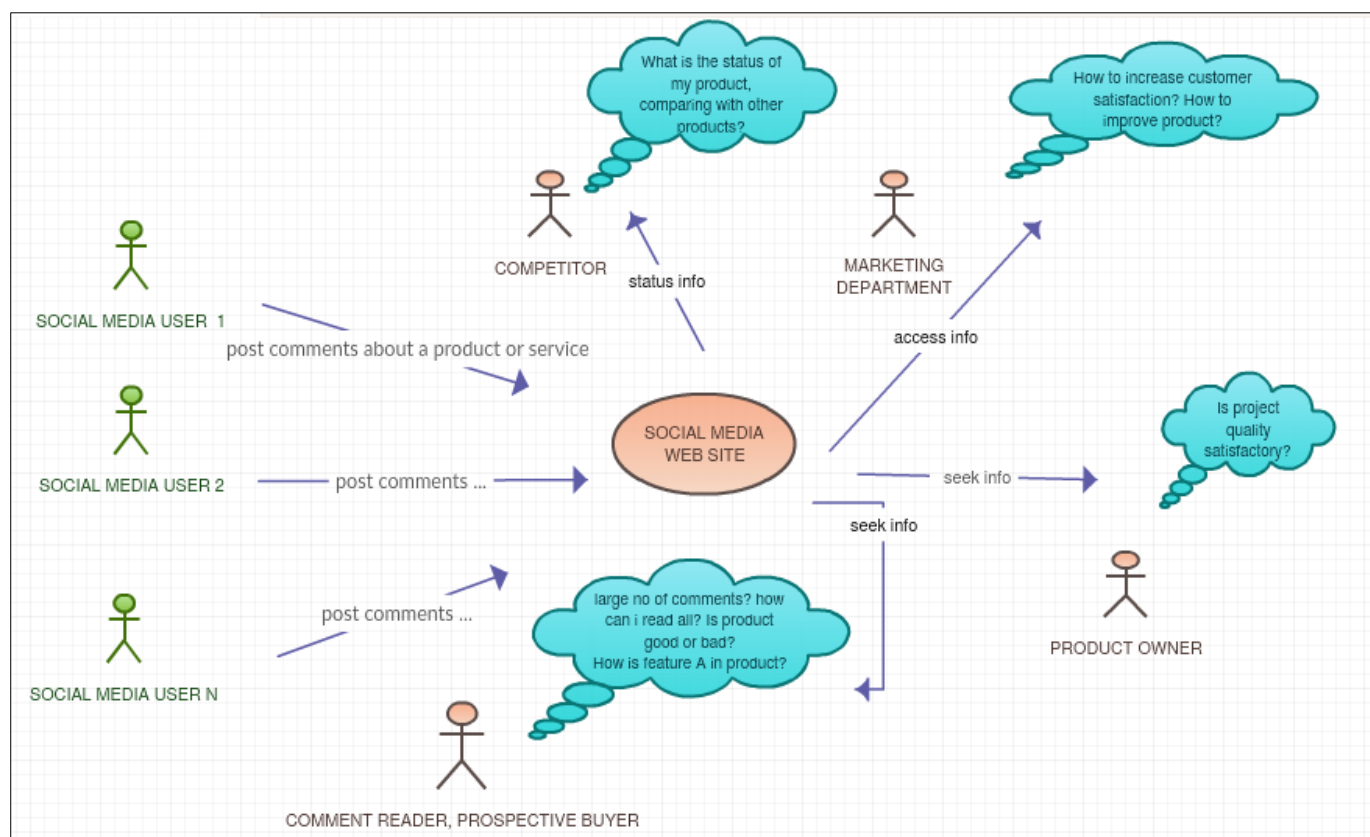


Figure 11: Rich Picture of OM Problem

4.3. Context Diagram

Following is the context diagram of OM application.

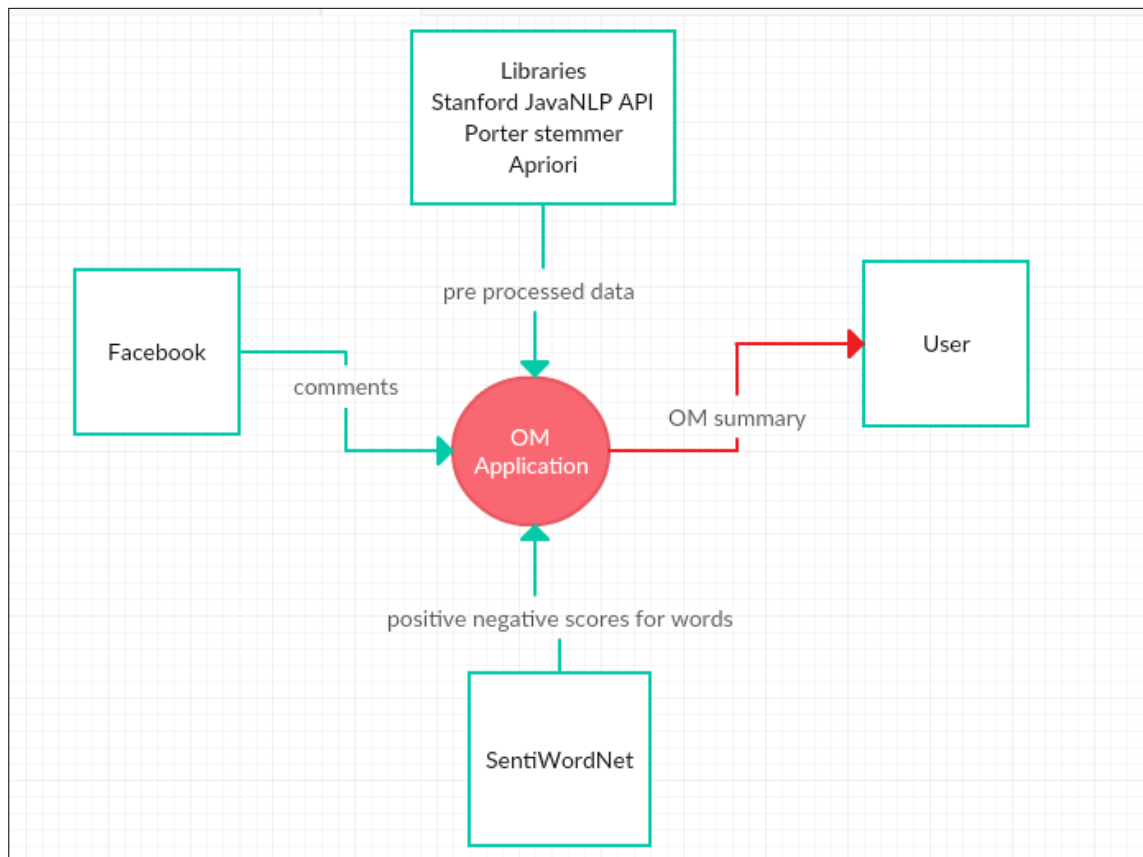


Figure 12: Context Diagram of OM Application

OM application connects to third party social media application Facebook via Graph API and retrieve comments, which is the input for the system. Then those comments are preprocessed using third party APIs like Stanford JavaNLP etc. and comments are processed with the help of SentiWordNet resource. Finally sentence level and aspect level summary is given to the end users for decision making.

4.4. Analysis

4.4.1. Stakeholder Analysis

Stakeholder analysis in conflict resolution, project management, and business administration, is the process of identifying the individuals or groups that are likely to affect or be affected by a proposed action, and sorting them according to their impact on the action and the impact the action will have on them.

Following is the stakeholder analysis for proposed system.

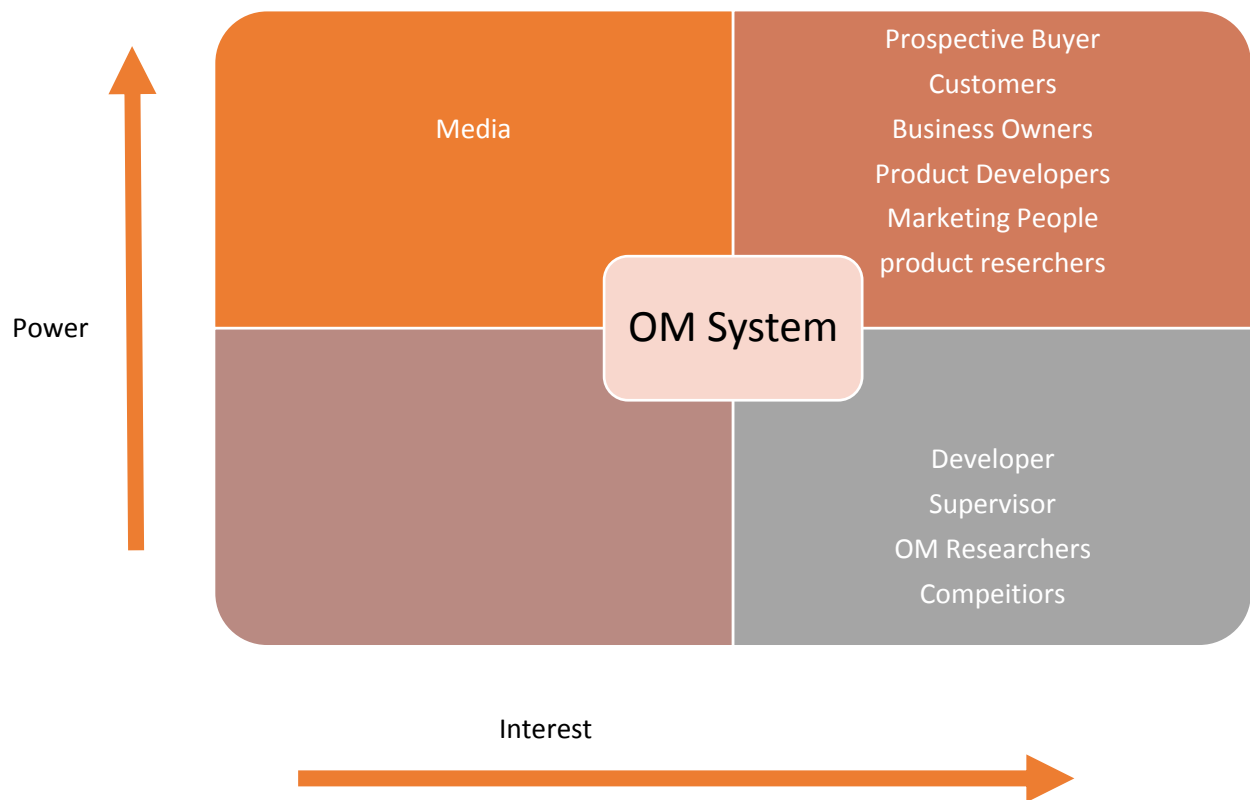


Figure 13: Power /Interest Grid for Stakeholder Prioritization

- High power, interested people - people need to manage closely and put greatest effort to satisfy by the system
- High power, less interested people – people need to keep informed
- Low power, interested people – people need to keep satisfied
- Low power, less interested people – people need to keep monitor with minimum effort

4.4.2. Requirement Analysis

Requirements of the project are identified by going through the research papers, current implemented systems and online articles about the OM process and systems.

4.4.2.1. Functional Requirements

Functional requirements describe what the system should do. In Software engineering and systems engineering, a functional requirement defines a function of a system or its component. A function is described as a set of inputs, the behavior, and outputs.

- FR - 1 : System should retrieve fb post comments via Graph API
- FR – 2 : System should store retrieved fb post comments in a text file

- FR – 3: System should preprocess the post comments
- FR – 4: System should extract aspects
- FR – 5: System should find opinion words
- FR – 6: System should calculate positivity negativity scores for the aspects
- FR – 7: System should calculate sentence scores
- FR – 8: System should identify the aspect positivity or negativity
- FR – 9: System should allow user to login
- FR – 10: System should allow user to provide the post id
- FR – 11: System should allow user to configure OM variables
- FR – 12: System should allow user to view the final report
- FR – 13: System should allow user to logout
- FR – 14: System should give a proper error message for invalid login
- FR – 15: System should give a proper message when user logout
- FR – 16: System should give a proper error message for invalid fb post id

4.4.2.2. Non Functional Requirements

Nonfunctional requirements describe how the system will do. In systems engineering and requirements engineering, a non-functional requirement is a requirement that specifies criteria that can be used to judge the operation of a system, rather than specific behaviors. They are contrasted with functional requirements that define specific behavior or functions.

- NFR – 1: At least 60% accuracy need to be gain from the system
- NFR – 2: System should be accessible consistently via web
- NFR – 3: System should be able to process at least 1000 comments at a time
- NFR – 4: System should calculate the scores and give the results in a user bearable time
- NFR – 5: System should be able to test via unit tests
- NFR – 6: System should be able to recover from errors
- NFR – 7: System should be able to extensible in the future by adding features or changing features
- NFR – 8: System should be user friendly for the target audience
- NFR – 9: System should be easy to operate
- NFR – 10: System should provide help information

4.4.3. Use Case Model

A use case is a list of actions or event steps, typically defining the interactions between a role (known in the Unified Modeling Language as an actor) and a system, to achieve a goal. The actor can be a human or other external system.

Following are the use cases of proposed system.

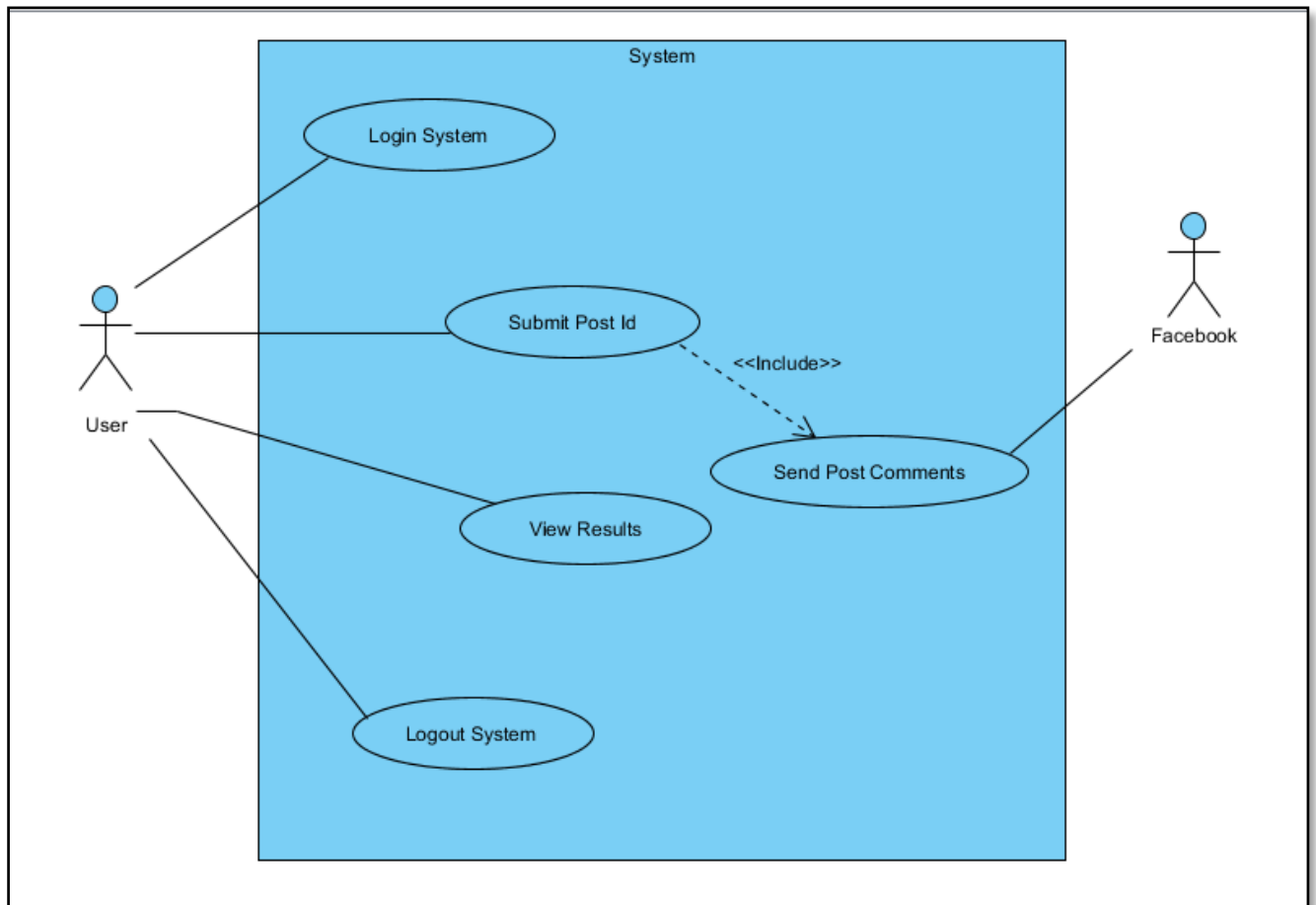


Figure 14: Use Case Diagram

- UC - 01: Login System
- UC – 02: Submit Post Id
- UC – 03: View Results
- UC – 04: Logout System
- UC – 05: Send Post Comments

4.4.4. Use Case Description

Table 1: UC-01: Login System

Use Case ID	UC - 01
Use Case Name	Login System
Description	Describes how to login the system
Pre-Condition	User should have a valid username and password
Post Condition	User should login the system
Actor	User
Main Flow	<ol style="list-style-type: none"> 1. Enter valid username and password 2. Click on Login button 3. System validates the credentials and user should login the system
Alternative Flow	-
Exceptions	<ol style="list-style-type: none"> 1. Enter invalid username or password 2. Click on Login button 3. System validates the credentials and shows the error message "Invalid Username or Password" <ol style="list-style-type: none"> 1. Enter the password 2. Click on Login button 3. System should show the error message "Username is required" <ol style="list-style-type: none"> 1. Enter the username 2. Click on Login button 3. System should show the error message "Password is required"

Table 2: UC-02: Submit Post Id

Use Case ID	UC - 02
Use Case Name	Submit Post Id
Description	Describe how to put the post id that need to be opinion minded
Pre-Condition	User should have login the system
Post Condition	User should see the processing image and finally should see the opinion mining results
Actor	User
Main Flow	<ol style="list-style-type: none"> 1. Enter valid post id in the text box 2. Click on Start button 3. System shows the processing image 4. System should direct user to the result page 5. User should be shown the results
Alternative Flow	-
Exceptions	<ol style="list-style-type: none"> 1. Click on Start button without entering a post id 2. System should show an error message "Post Id Cannot be empty." <ol style="list-style-type: none"> 1. Enter an invalid post id 2. Click on Start button 3. System should show an error message "Error occurred while retrieving the comments."

Table 3: UC-03: View Results

Use Case ID	UC - 03
Use Case Name	View Results
Description	Describe how results are shown to the user
Pre-Condition	User should have login the system
Post Condition	User should see the aspects results, sentence results and summarization results
Actor	User
Main Flow	<ol style="list-style-type: none"> 1. Enter a valid post id 2. Click on Start button 3. System should show the results to the user
Alternative Flow	-
Exceptions	-

Table 4: UC-04: Logout System

Use Case ID	UC - 04
Use Case Name	Logout System
Description	Describe how user logout from the system
Pre-Condition	User should have login the system
Post Condition	User should logout from the system
Actor	User
Main Flow	<ol style="list-style-type: none"> 1. Click on logout button 2. User should logout from the system 3. User should prompt the Login page 4. "You've been logged out successfully." message should be shown in the login page
Alternative Flow	-
Exceptions	-

Table 5: UC-05: Send Post Comments

Use Case ID	UC - 05
Use Case Name	Send Post Comments
Description	Describe how Facebook send post comments to the OM system
Pre-Condition	User should have login the system Facebook user access token should be valid
Post Condition	OM system should receive post comments from Facebook
Actor	Facebook
Main Flow	<ol style="list-style-type: none"> 1. Enter post id in Home page 2. Click on Start button 3. System should send a POST request to Facebook via Graph API 4. Facebook should send comments for the post id via POST body 5. OM system should start preprocessing the post comments
Alternative Flow	-
Exceptions	<p>With invalid Facebook user access token.</p> <ol style="list-style-type: none"> 1. Enter post id in Home page 2. Click on Start button

	<ol style="list-style-type: none">3. System should send a POST request to Facebook via Graph API4. Facebook should send an error to OM system5. System should show an error message “Error occurred while retrieving the comments.”
--	---

4.5. Summary

This chapter described about the analysis done on stakeholders and requirements. In the requirement analysis section functional and non-functional requirements of the system are identified and presented. Later the use case diagram is presented describing the requirements in user perspective.

5. System Design and Architecture

Introduction

Proposed System

High Level Design of the System

System Architecture

Flow Charts

Class Diagrams

Sequence Diagrams

Summary

5.1 Introduction

This chapter describes the proposed system design and architecture. Software design is the process of transforming user requirements into form, which helps developers to implement the system. Following is a brief description of system architecture and design.

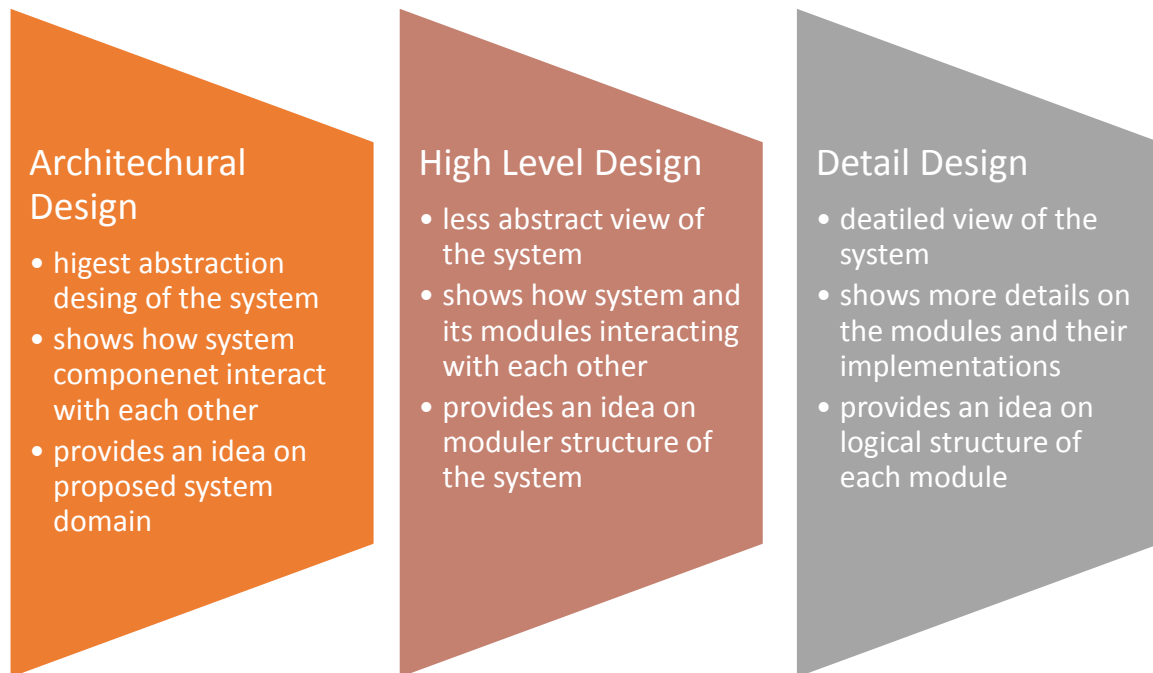


Figure 15: System Architecture and Design

5.2 Proposed OM System

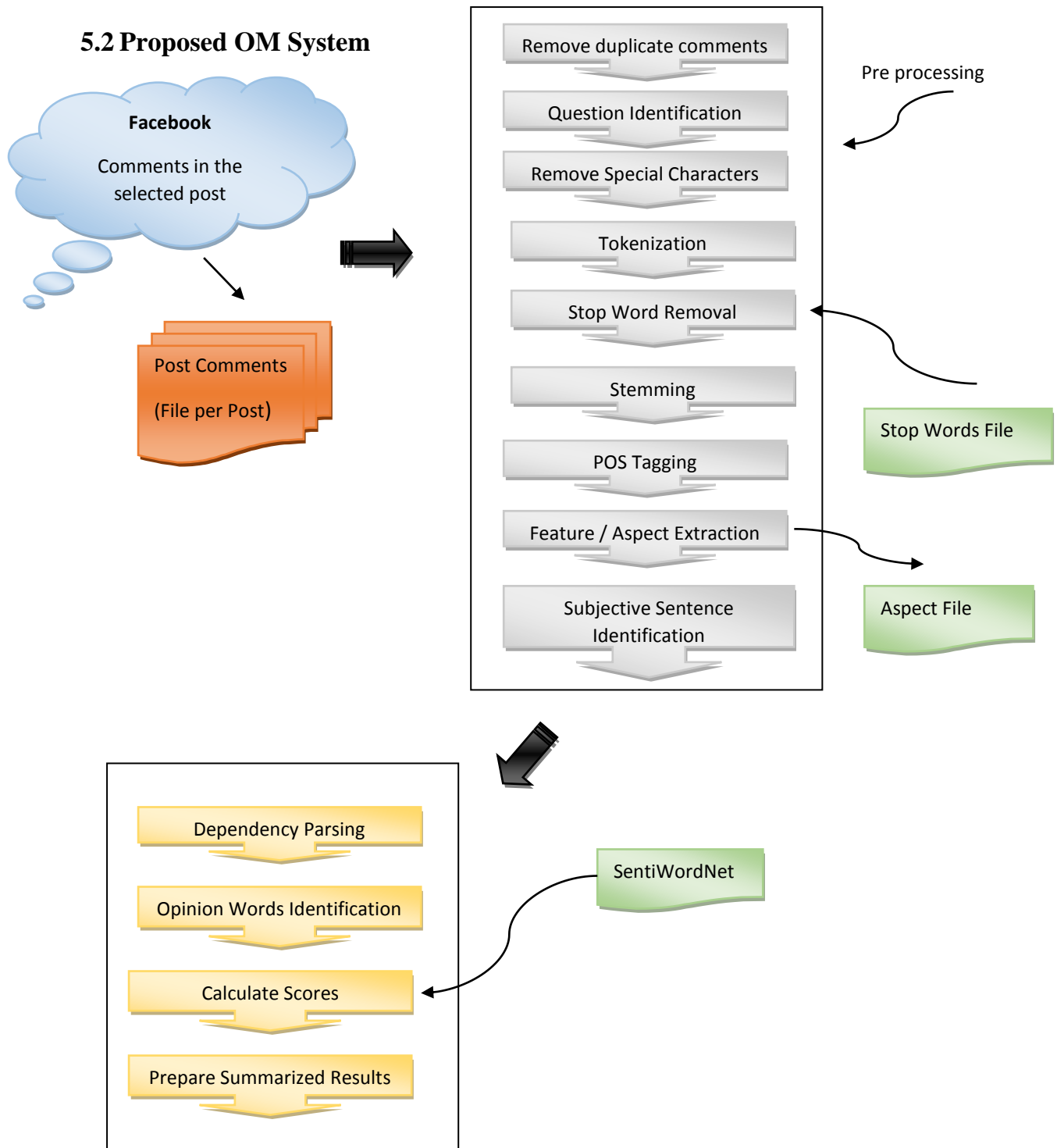


Figure 16: Opinion Mining System Overview

Above diagram shows the proposed OM system. Facebook comments are retrieved via Graph API, using user access token and saved them in separate files per post. Then preprocessing steps are carried out with the help of pre saved stop words file and pre-configured configuration values. After cleaning process is done opinion mining process is started by identify the grammatical structure of the comments by dependency parsing. Then opinion words are identified and scores are calculated

for discovered aspects and sentences. Finally, results are summarized and presented to the users in a user friendly way.

5.3 High Level Design of the Opinion Mining System

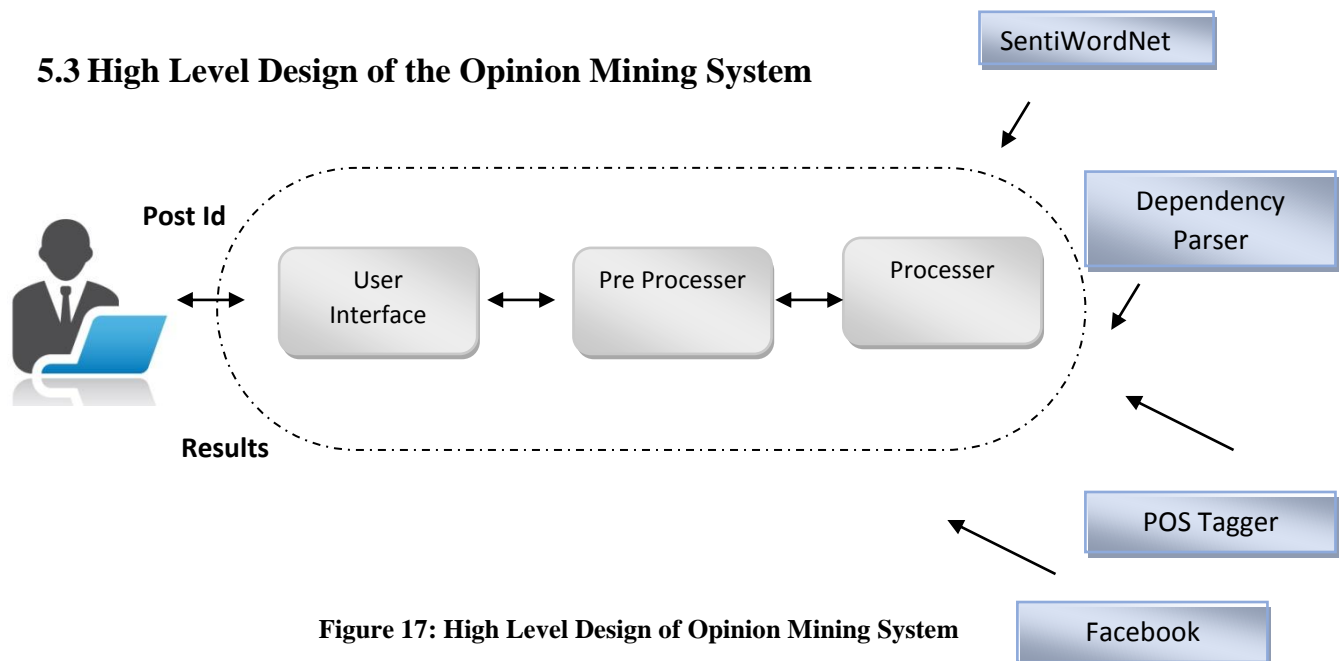


Figure 17: High Level Design of Opinion Mining System

Opinion mining system is a syntactic based application. User interacts with the user interface and backend components interact with Facebook, Stanford Dependency Parser, Maxent Tagger and SentiWordNet lexicon.

5.4. System Architecture

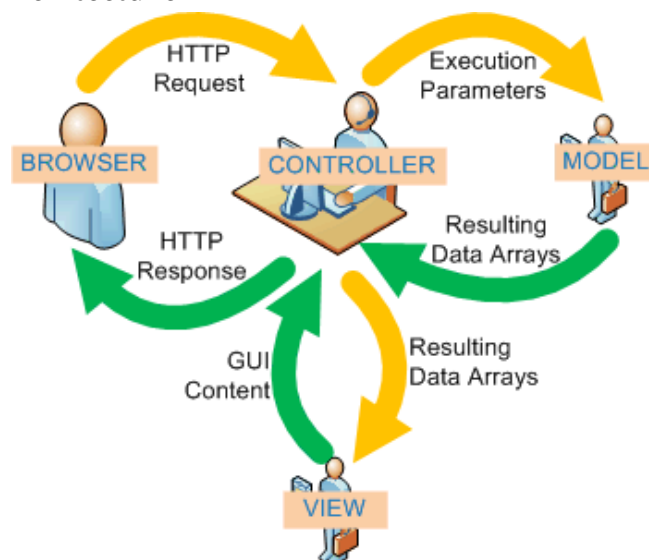


Figure 18: System Architecture

Opinion mining system is designed based on MVC architecture. MVC is a popular design pattern used today and it isolates the user interface layer with backend layer. All the incoming requests are received to jsp pages and they transferred the incoming requests to controllers using Ajax and JavaScript technologies. Then controllers delegate the requests to appropriate services. Services

performed the business logic such as preprocessing, processing etc. according to the logics and business rules defined in the classes and produce relevant results. Finally, the results are sent back to the User Interface layer.

Here the system does not incorporate with any database system. The text files are used to save relevant results, which are needed to access in future processes. Therefore, the data files act as the model in this architecture. Since file read write operations are costly, special attention has been given to minimize the file read write operations and increase the efficiency, performance as much as possible.

In addition, the system has externalized the main property file, which has configurations like log configurations, Facebook access details, Apriori algorithm properties, file save paths etc. to minimize the dependency between systems and make easy for the user to set the values as their preferences.

5.5. Data Flow Diagram

DFD Data Flow Diagram represents the data flow like incoming data, outgoing data, store data and processes of the system. Following is the high level DFD diagram of Opinion Mining Application.

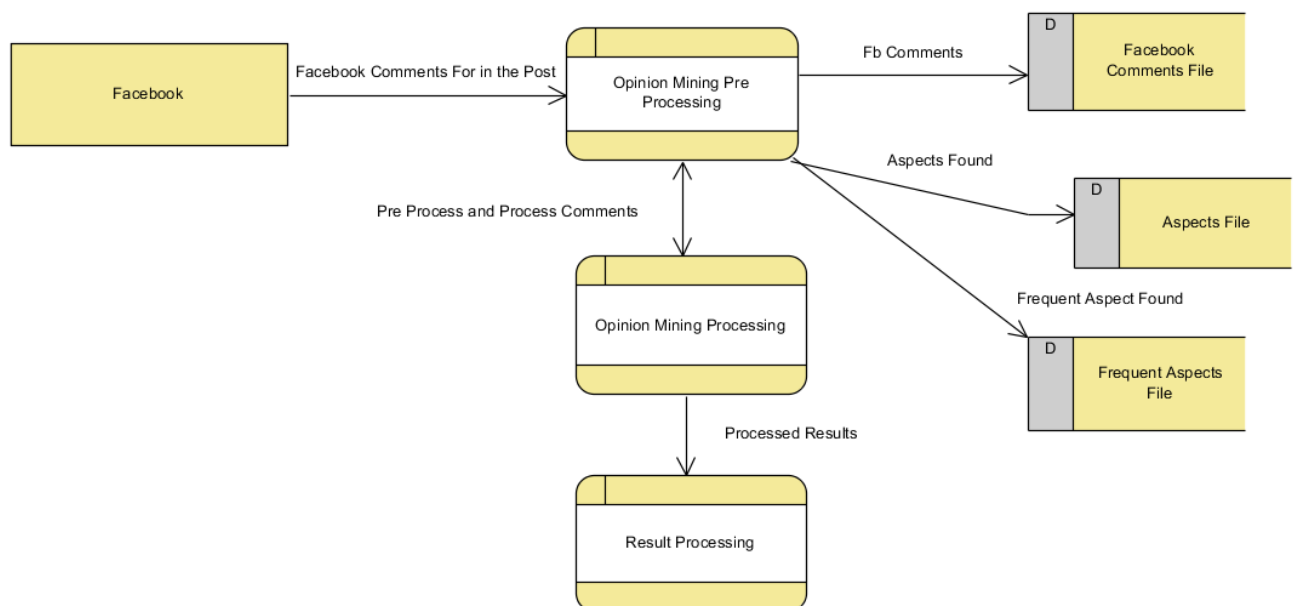


Figure 19: Data Flow Diagram

5.6. Class Diagram

Class diagram presents the static view of the system. It shows the classes and their relationship.

User Login

In user Login flow login.jsp acts as the View, MainController.java acts as the controller and WebSecurityConfig.java will acts as the model as it is configured with username and passwords for the valid users. Following table lists MVC components (java classes, jsp pages) in each flow of the system.

Module Name	Model	View	Controller
Login	WebSecurityConfig.java	login.jsp Home.jsp	MainController.java
Log Out	WebSecurityConfig.java	logout.jsp	-
Retrieve Comments	Comment.java FbComment.java From.java	Home.jsp	OpinionMiningController.java
Opinion Mining Pre Processor	comments stopwords removewords aspects frequent aspects	home.jsp	OpinionMiningController.java OpinionMiningPreProcessingService.java OpinionMiningPreProcessingServiceImpl.java FileWriter.java PorterStemmer.java OMAppConfiguration.java FileLoader.java MaxentTagger.java AlgoApriori.java
Opinion Mining Processor	processed comments frequent aspects results		OpinionMiningProcessingService.java OpinionMiningProcessingServiceImpl.java OMAppConfiguration.java SentiWordNetScoreCalculator.java
Opinion Mining Result Processor	AspectResult OpinionMiningProcessingResult.java Polarity.java SentenceResult.java	result.jsp	OpinionMiningController.java

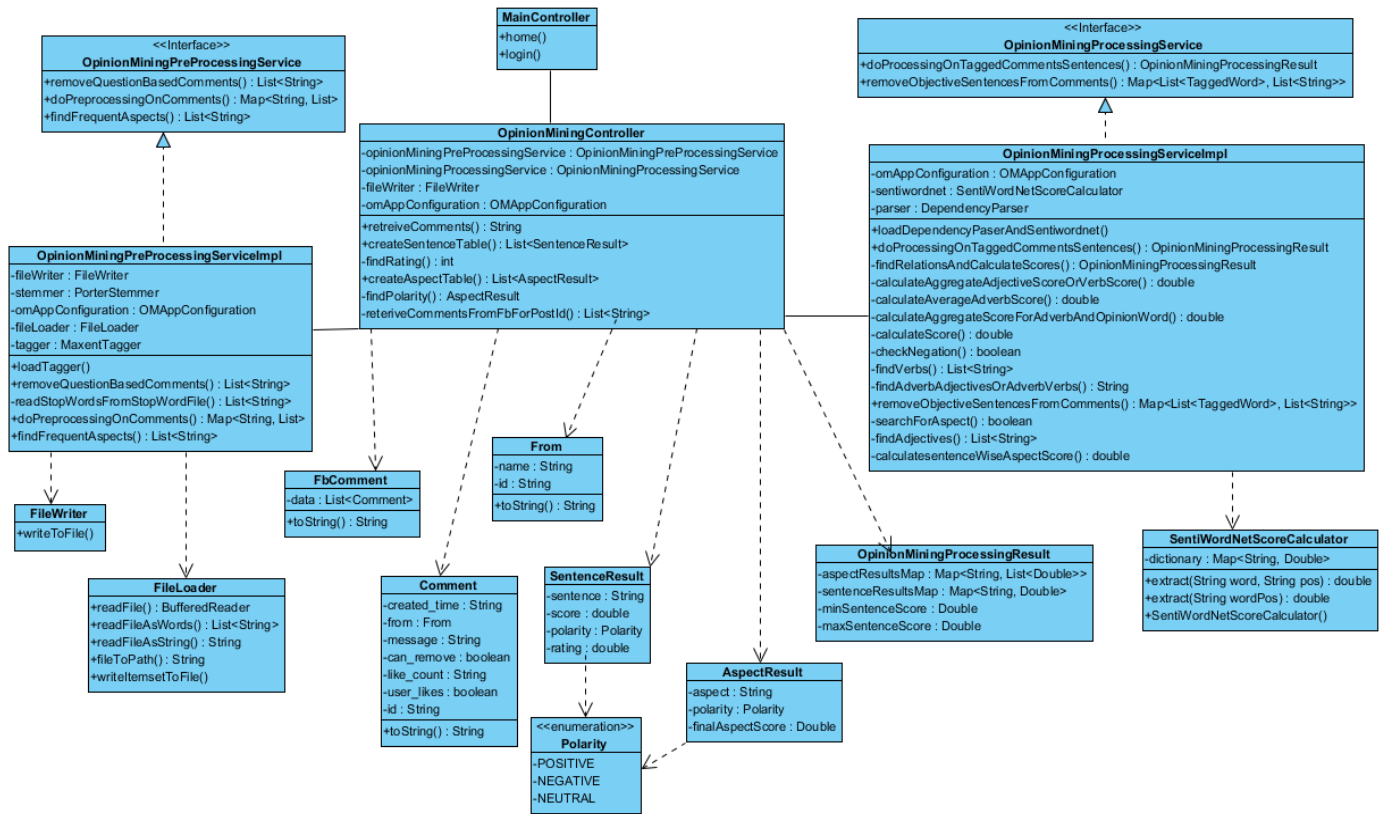


Figure 20: Class Diagram for System

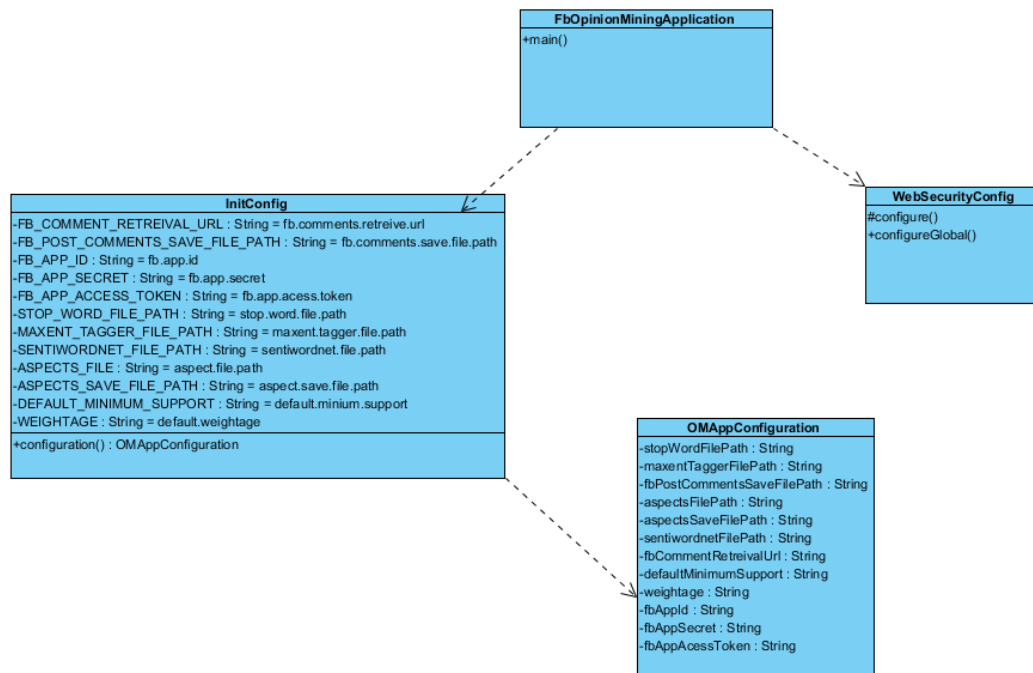


Figure 21: Class Diagram for System Loading

5.7. Sequence Diagrams

A Sequence diagram is an interaction diagram that shows how objects operate with one another and in what order. It is a construct of a message sequence chart. Following are the sequence diagrams of OM system.

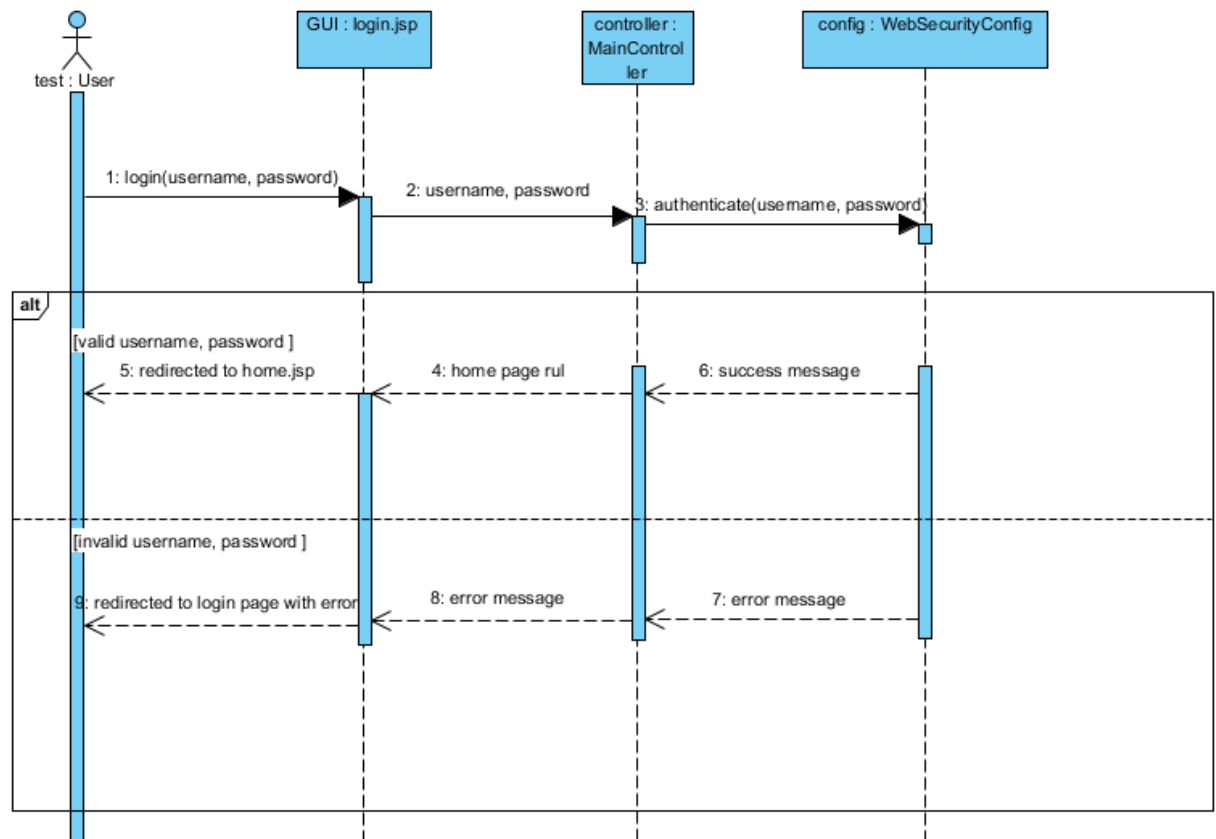


Figure 22: Sequence Diagram for Login

The diagram shows the login process of the system. User entered username and password is validated against the configured users in memory authentication process. Valid users are redirected to the home.jsp and invalid users are redirected to the login page with an error message.

The below diagram shows the process of retrieving comments from Facebook. Once the user logged in to the system, user enters the post id which opinion mining should be done. System sends a POST request to Facebook Graph API to get the posts from Facebook. If the access token and post id is correct Facebook sends the posts comments in JSON format. Otherwise, an exception is thrown from Facebook and OM system handles the exception and user friendly error message is shown to the end user.

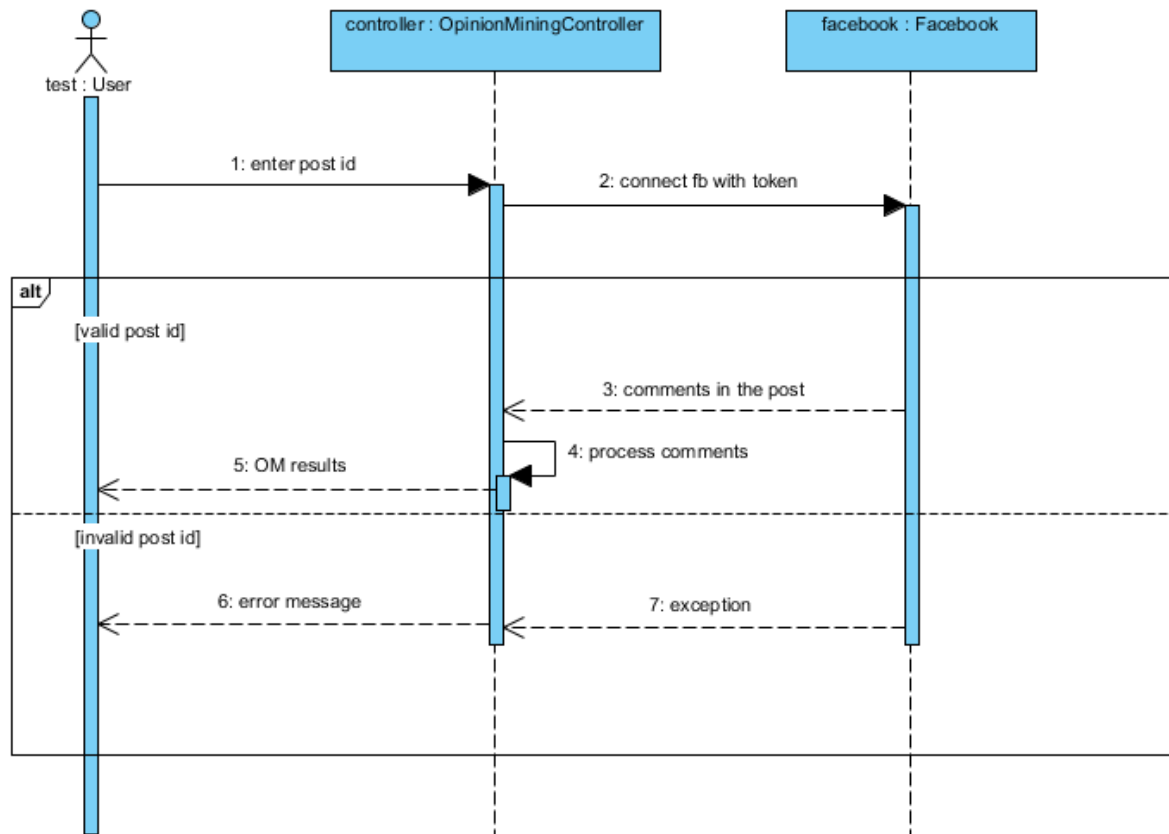


Figure 23: Sequence Diagram for Retrieve Facebook Comments

Below diagram shows the sequences of preprocess flow. Once the comments are retrieved from Facebook, OM system preprocesses the comments using below steps.

- Filter comments by author id and remove duplicate author comments to avoid opinion spamming.
- Remove question based comments (sentences) to avoid question based comments.
- Write filtered comments to a file to future reference.
- Tokenize sentences using Maxent Tagger.
- Load stop words file from file system.
- Remove stop words from comments by comparing the comment with stop words in loaded stop words file
- Stem the sentences
- Tag the sentences
- Find aspects in the sentences
- Write found aspects to a file for future reference
- Find frequent aspects using Apriori algorithm
- Finally send preprocessed comments to OpinionMining Controller

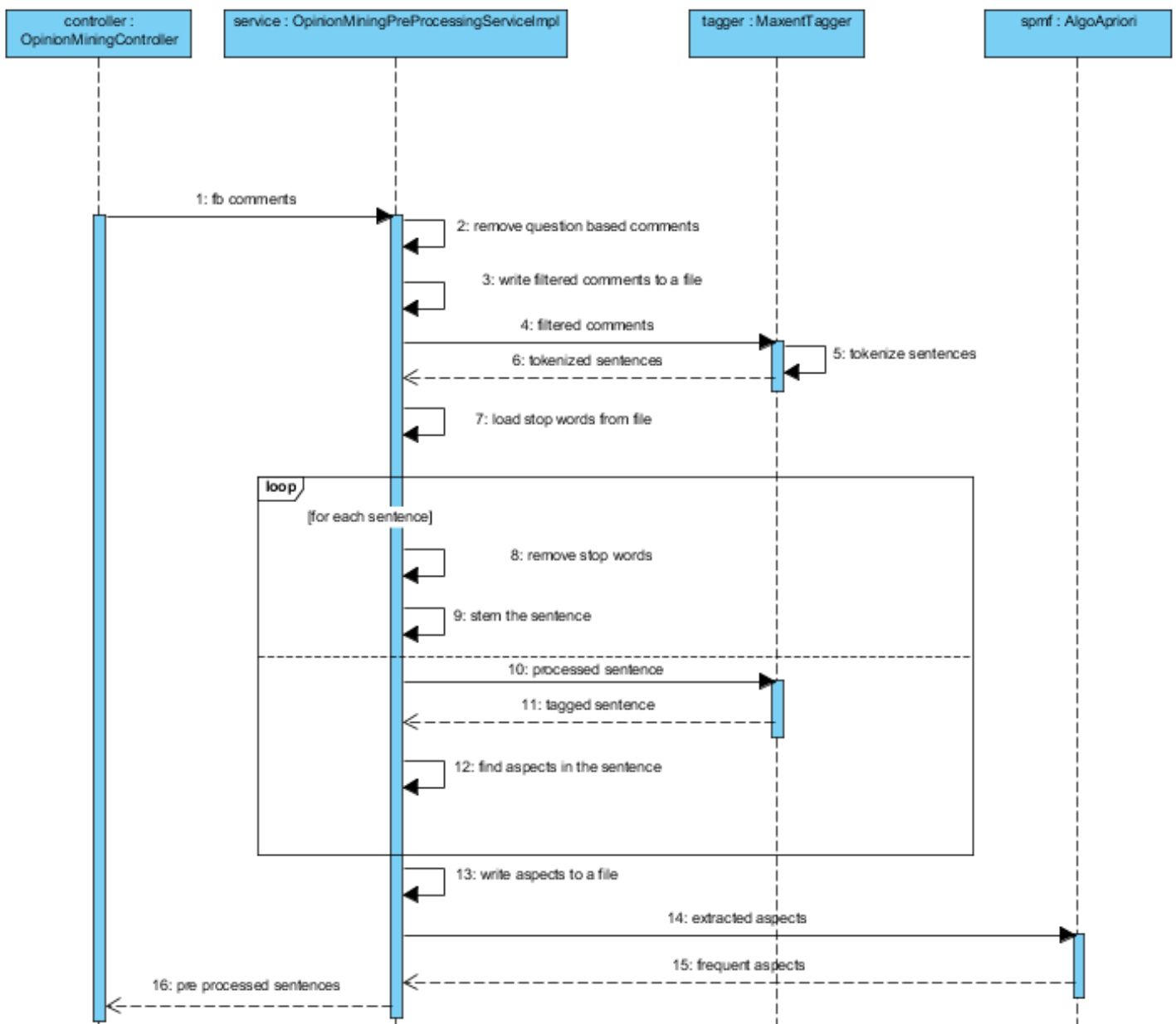


Figure 24: Sequence Diagram for Pre Process Comments

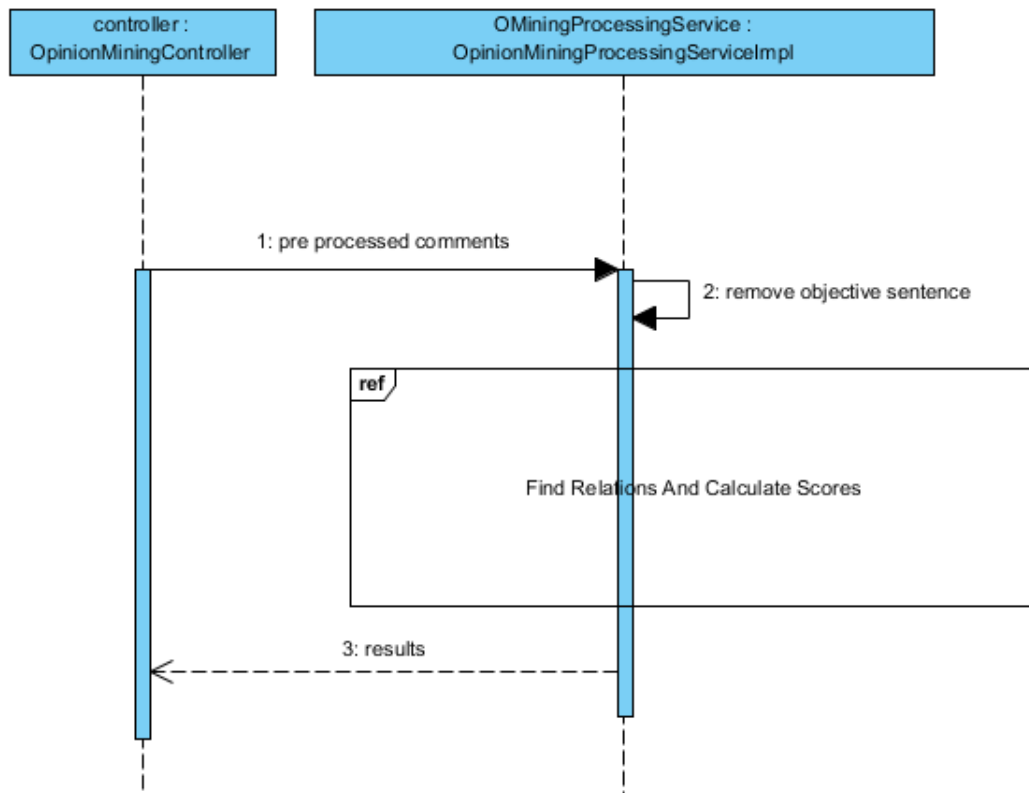


Figure 25: Sequence Diagram for Process Comments

Above diagram shows the process of processing comments. Preprocessed comments are sent to OpinionMiningProcessingImpl service class. It filters the subjective sentence by removing objective sentences from the comments. (Subjective sentence is a sentence, which contains at least a frequent aspect.) Then the system finds relations in the sentence and calculates scores for opinion words. Following diagram shows the detail flow of process.

- Load required dependency parser model – Here default model – “**english_UD.gz**” is used as the parser model
- Parse the sentences using parser model and identify the grammatical structure of the sentence
- Then for each aspect in sentence, continue the following steps
 - Find adjectives
 - Calculate aggregate adjective score
 - Find verbs
 - Calculate aggregate verb score
 - Calculate sentence wise aspect score
 - Calculate the sentence score

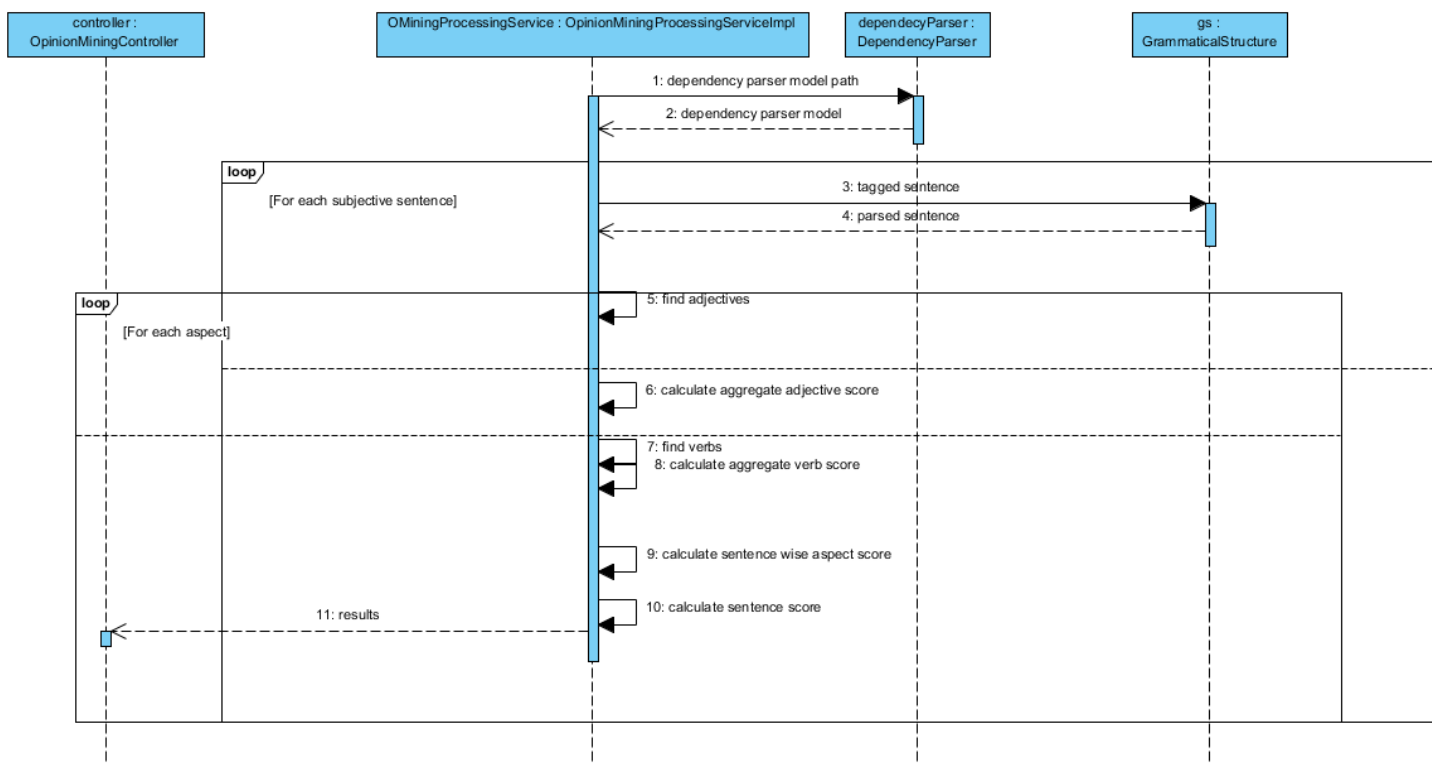


Figure 26: Sequence Diagram for Find Relations and Calculate Scores

5.8. Summary

Requirements are identified in the previous chapter and this chapter has described how the system is designed according to them. Firstly, overview of the system is drawn. Then high level design and system architecture are described using diagrams. Class diagram is presented to show the interactions between classes and data flow diagram is drawn to visualize how data of the system is coming in and going out. Finally, sequences diagrams are presented to each process to get a detail idea on how interactions are happened between objects of the system.

6. Implementation

Introduction

Comments Retrieval

Pre Processing

Processing

SentiWordNet

Stanford POS Tagger

Apriori Algorithm

Summary

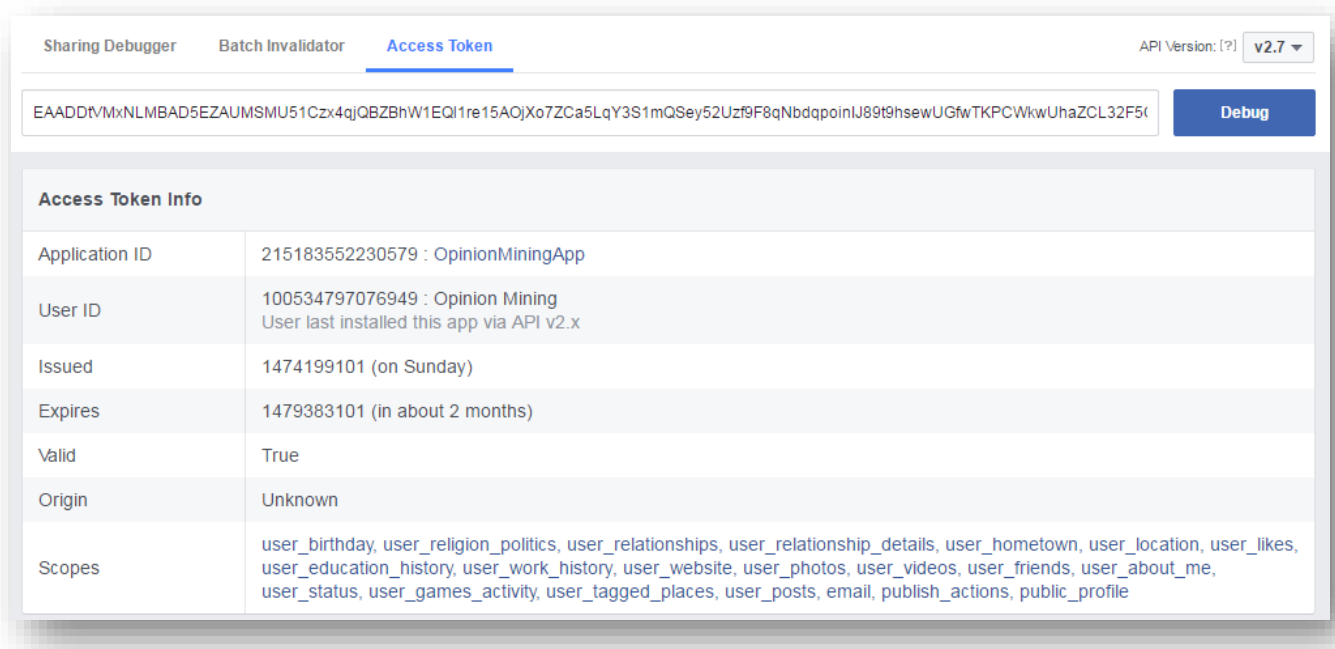
6.1. Introduction

This chapter gives a detail description of the implementation which has been done based on the design described in previous chapter. Here all the functions of the system are described along with the tools, algorithms, frameworks, technologies, resources and libraries used.

6.2. Comments Retrieval

In the system Facebook post comments are retrieved and send for the opinion mining process. Graph API is used to retrieve comments and the relevant profile that comments are being taken need to give the access to Graph API to access the profile. Or else comments can be retrieved via a Facebook application but the relevant Facebook application should be given the rights to access the profile or page. Access is given using a user token generated via Facebook token tool.

Here the post comments are retrieved from Facebook profile “Opinion Mining.” post id, user id and user access key is sent along with the following query to be run on Graph API.



The screenshot shows the Facebook Access Token tool interface. At the top, there are tabs for 'Sharing Debugger', 'Batch Invalidator', and 'Access Token'. The 'Access Token' tab is selected. Below the tabs, there is a text input field containing a long alphanumeric string (the access token) and a 'Debug' button. Below this, there is a section titled 'Access Token Info' which displays the following details:

Application ID	215183552230579 : OpinionMiningApp
User ID	100534797076949 : Opinion Mining User last installed this app via API v2.x
Issued	1474199101 (on Sunday)
Expires	1479383101 (in about 2 months)
Valid	True
Origin	Unknown
Scopes	user_birthday, user_religion_politics, user_relationships, user_relationship_details, user_hometown, user_location, user_likes, user_education_history, user_work_history, user_website, user_photos, user_videos, user_friends, user_about_me, user_status, user_games_activity, user_tagged_places, user_posts, email, publish_actions, public_profile

Figure 27: Facebook User Access Token

User Access Token

The user token is the most commonly used type of token. This kind of access token is needed any time the app calls an API to read, modify or write a specific person's Facebook data on their behalf. User access tokens are generally obtained via a login dialog and require a person to permit your app to obtain one. (facebook for developers, 2016)

Sample User Access Token

```
EAADDtVMxNLMBAAkFAZADrZBFIZCKWkUdNHjh0wLG6ugB0Hz4StvyCw92e5jqtd7w79njhikVfusFqo8YpMhb1kiqD68IKcZCiFEGGnjMFlemz3aJSbggLxWoA8fwG1TtiLOje2ifNVC5zAYYfFb1gpO644ahRcZD
```

Query to Retrieve Comments from Facebook Post

https://graph.facebook.com/v2.7/{post_id}/comments?

{post_id} → (fb user id)_(fb post id)

Example post id = 100013612860254_105490496581379

Response JSON retrieved from Facebook

```
{
  "data": [
    {
      "created_time": "2016-09-22T05:43:34+0000",
      "from": {
        "name": "Opinion Mining",
        "id": "100534797076949"
      },
      "message": "But if I installed either one of these Norton products, neither works after installation?",
      "id": "105491009914661_105491429914619"
    },
    {
      "created_time": "2016-09-22T05:43:47+0000",
      "from": {
        "name": "Opinion Mining",
        "id": "100534797076949"
      },
      "message": "It can not be the computer or the owner, since I purchased McAfee Anti-Virus 8 and it installs and works fine with no problems.",
      "id": "105491009914661_105491743247921"
    },
    {
      "created_time": "2016-09-22T05:43:59+0000",
      "from": {
        "name": "Opinion Mining",
        "id": "100534797076949"
      },
      "message": "I have used Norton products in the past and I am familiar with them.",
      "id": "105491009914661_105491813247914"
    },
    {
      "created_time": "2016-09-22T05:44:06+0000",
      "from": {
        "name": "Opinion Mining",
        "id": "100534797076949"
      },
      "message": "I bought NIS 2004 recently to try it out.",
      "id": "105491009914661_105491943247901"
    }
  ],
  "paging": {
    "cursors": {
      "before": "WTI5dGJXVnVkrJlqZAFhKemIzSTZANVEExTkRreE1UQTJOVGd4TXpFNE9qRTB0e1ExTWpJNU9EQT0ZD",
      "after": "WTI5dGJXVnVkrJlqZAFhKemIzSTZANVEExTkRrM01qYzVPVEUwTURNME9qRTB0e1ExTWpNM09Eaz0ZD"
    }
  }
}
```

JSON

JavaScript Object Notation) is a lightweight data-interchange format. It is widely used in parse data between servers. Also, it is in human readable format. Facebook comments are retrieved as a JSON array and OM application casts it to “FbComment” domain object using Jackson library which is a popular and efficient java based library which converts JSON to java objects and vice versa. To send and retrieve comments REST API is used.

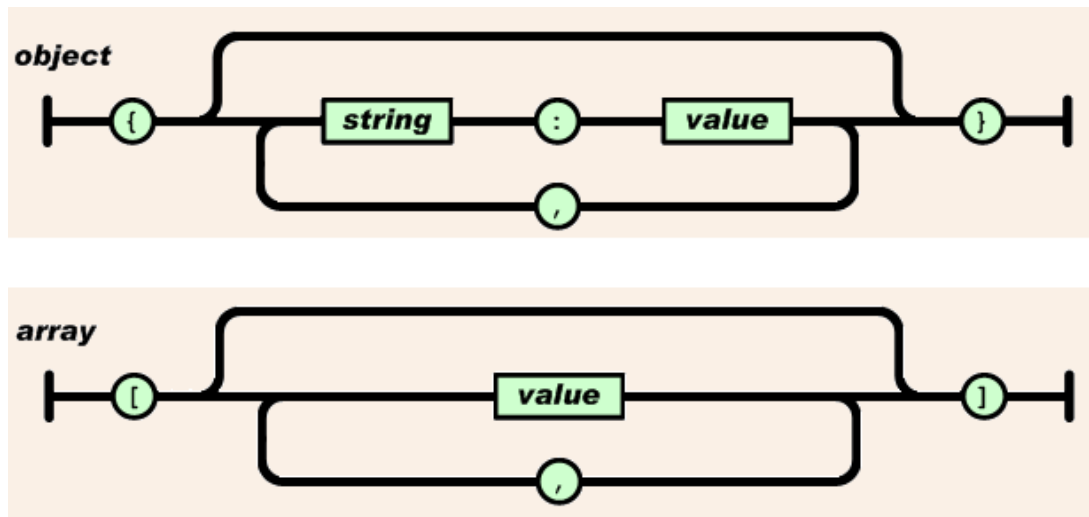


Figure 28: JSON Object and Array

REST

REST - Representational State Transfer uses http protocol to crud (Create/Read/Update/Delete) operations. In the system following rest operations are used,

Request Mapping	Request Method	Explanation
/, /home	GET	Load home page
/login	GET	Load login page
/retreivecomments	POST	Connect Facebook to retrieve post comments

Table 6: REST Operations

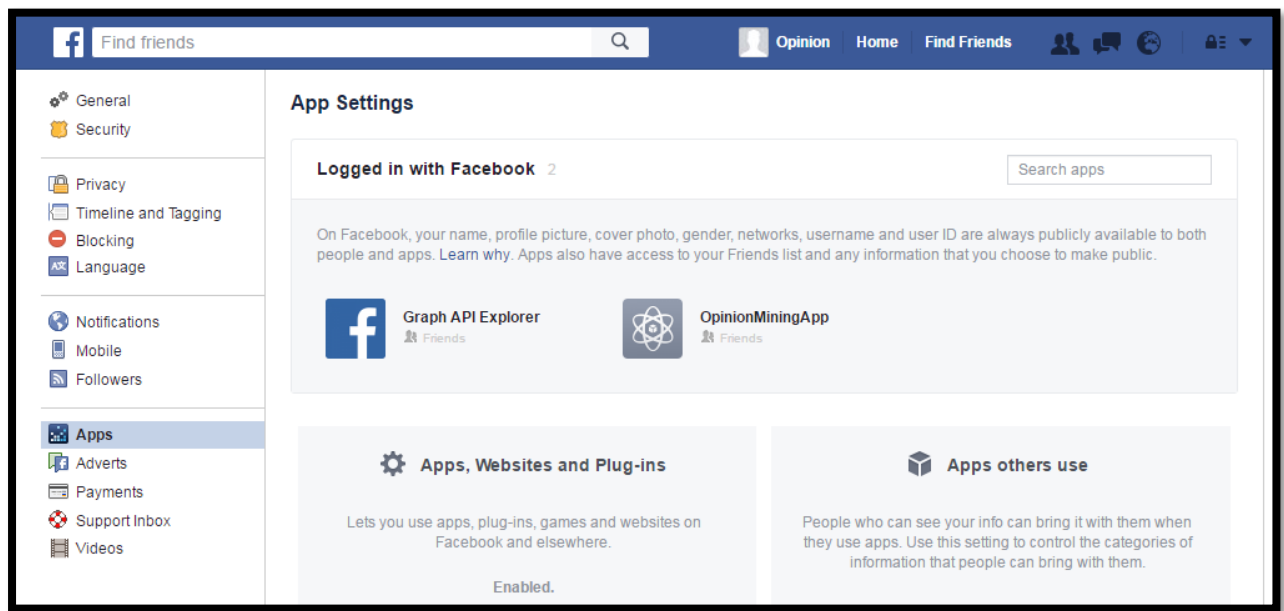


Figure 29: Facebook App Settings

“spring-boot-starter-social-facebook” spring based library is used to connect fb platform and access fb data. It provides simplest methods to connect and fetch fb data.

Graph API is the easiest and main method to communicate with Facebook platform. It has several versions and contains operations like query data, post stories etc. Here version 2.8 is used to communicate with Facebook.

6.3. Pre Processing

Preprocessing is the process done before OM, which removes the unnecessary processing overhead. In addition, it improves the accuracy and efficiency of OM process.

6.3.1. Remove Duplicate Author Comments

In the comments, there can be comments from a same user. It can be due to opinion spamming or due to a mistake. Therefore, those comments are filtered and only one comment will be taken in to next processes. Comments are filtered via “From → Id” in response JSON.

6.3.2. Remove Question Based Comments

In the comments there can be questions posted by the reviewers or users. They cannot be categorized as positive or negative thus need to be filter out and removed before aspect

identification. Most of the question based sentences are ended up with “?” mark. Therefore, as a preprocessing step sentences which are ended up with “?” will be removed and saved in a text file for later analytical process.

6.3.3. Case Normalization

Turn entire document to lowercase or uppercase. It will reduce the processing overhead of next steps.

6.3.4. Comments Tokenization

Tokenization is the process of chopping the sentence in to pieces, which are called as tokens. They can be referred as terms or words in the usage. Tokenization will reduce the processing overhead in the next phases and increase the efficiency of the process.

For the tokenization Stanford JavaNLP API – MaxentTagger will be used. Stanford JavaNLP API is a library provided by The Stanford NLP Group.

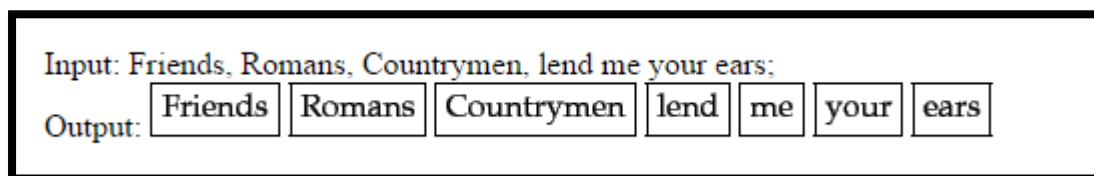


Figure 30: Tokenization Sample

The Stanford NLP Group

The Stanford NLP Group is a natural language group of Stanford University who makes natural language software available to everyone. Group is consisting of post docs, programmers and students who works on algorithms, which allows computers to understand and process human languages.

They have published several software's like Stanford CoreNLP, Stanford Parser, Stanford POS Tagger, Stanford Named Entity Recognizer, Stanford RegexpNER, Stanford Word Segmenter, Stanford Classifier, Stanford EnglishTokenizer etc and they are widely used in industry, academia and government. Some of the software's like Stanford Parser, Tagger will be used in this project.

6.3.5. Remove Stop Words

Remove unwanted words like language specific functional words which carry no information in each review sentence is called stop word removal. Stop words can be pronouns, prepositions or conjunctions.

E.g.: - is, are, was etc.

Reviews are stored in a text file and they are given as the input to stop word removal. In addition, stop words are collected and stored in a text file and stop words are removed by checking against the stop words text file.

Stop word is an often heard phrase in opinion mining. Stop words are the commonly used words in a language and stop words removal helps to increase the efficiency in an opinion mining application since it helps to focus on the most important words without considering the commonly used words in a language.

Stop word removal is done as a preprocessing activity before parsing the text. Stop words can be categories in to different forms like determiners, coordinating conjunctions, prepositions etc. and removal of stop words depends on the need of the application. For some application, it can be helpful to remove all stop words while another application to be removed only determiners. In our prospective opinion mining system it will be enough to remove only the basic set of stop words like determiners - the, a, an, another since removing all can lead the application to miss the important words for mining.

The car is good.

Eg: - “good” can be a stop word in some scenario but in the opinion mining application it can be the important opinion about an aspect.

Some people have submitted stop word lists that can be used in public. Also several stop word removal tools can be found in the internet. Apache Lucene is the most popular free and open source informational retrieval software library that can be used to remove stop words using StandardAnalyzer and StopAnalyzer. Lucene provide lot of features in text mining. It has lot of releases and update frequent. The documentation available in the internet is bit old when compared with the latest release and bit complex to use. In addition, it is good if we use it for several analyzing methods and less efficient to use it only the for the stop word removal process. Therefore, simple stop words removal process is implemented using java in this system.

The default `StopAnalyzer.ENGLISH_STOP_WORDS_SET`

```
"a", "an", "and", "are", "as", "at", "be", "but", "by",
"for", "if", "in", "into", "is", "it",
"no", "not", "of", "on", "or", "such",
"that", "the", "their", "then", "there", "these",
"they", "this", "to", "was", "will", "with"
```


Here in this system, a custom stop word list with java implementation of stop word removal is used which is adequate for the purpose.

Stop word list used in this project

a	an	another	any	certain	near	next	of	off	on
each	every	her	his	my	onto	out	outside	over	past
no	our	some	that	the	plus	round	since	than	through
their	this	and	but	or	to	toward	under	underneath	unlike
yet	for	nor	so	as	until	up	upon	with	without
aboard	about	above	across	after					
against	along	around	at	before					
behind	below	beneath	beside	between					
beyond	but	by	down	during					
except	following	for	from	in					
inside	into	like	minus	opposite					

Table 7: Stop Words List

6.3.6. Stemming

Stemming is the process of forming root words of a word.

E.g.: - stemming algorithm reduces the words “longing”, “longed” and “longer” to the root word “long”.

There are lot of stemming algorithms like, n-gram analysis, Affix stemmers, Lemmatization algorithms etc.

Porter stemmer Algorithm is used to form root word from the given input reviews in this project.

Porter Stemmer Algorithm

Stemming is the process of removing the commoner morphological and in flexional endings from words in English. This is written by Martin Porter and algorithm is widely used in implementing information retrieval systems.

6.3.7. POS Tagging

Another important face of opining mining is to determine the features and opinion words. POS tagging is the process of labeling each word in a sentence with its appropriate part of speech – verb, noun, adjective, adverb, pronoun, preposition, conjunction, and interjection (linguistic category that

is defined by its syntactic or morphological behavior). This can be done manually or with the help of POS tagger tool. Stanford tagger – MaxentTagger is used in this system.

Part-of-Speech:

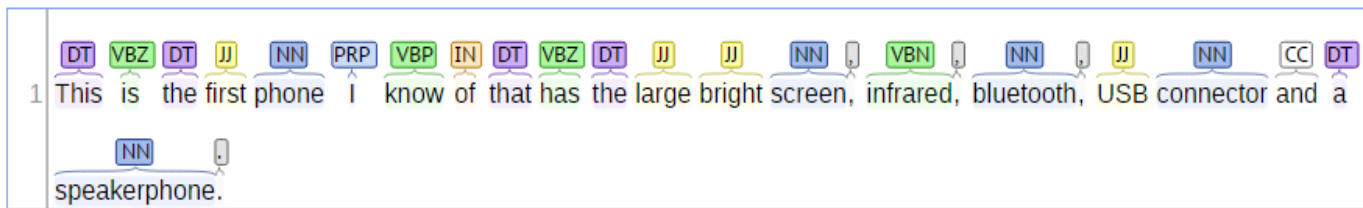


Figure 31: POS Example

Alphabetical list of part-of-speech tags used in the project

Number	Tag	Description	Number	Tag	Description
1.	CC	Coordinating conjunction	19.	PRP\$	Possessive pronoun
2.	CD	Cardinal number	20.	RB	Adverb
3.	DT	Determiner	21.	RBR	Adverb, comparative
4.	EX	Existential <i>there</i>	22.	RBS	Adverb, superlative
5.	FW	Foreign word	23.	RP	Particle
6.	IN	Preposition or subordinating conjunction	24.	SYM	Symbol
7.	JJ	Adjective	25.	TO	<i>to</i>
8.	JJR	Adjective, comparative	26.	UH	Interjection
9.	JJS	Adjective, superlative	27.	VB	Verb, base form
10.	LS	List item marker	28.	VBD	Verb, past tense
11.	MD	Modal	29.	VBG	Verb, gerund or present participle
12.	NN	Noun, singular or mass	30.	VBN	Verb, past participle
13.	NNS	Noun, plural	31.	VBP	Verb, non-3rd person singular present
14.	NNP	Proper noun, singular	32.	VBZ	Verb, 3rd person singular present
15.	NNPS	Proper noun, plural	33.	WDT	Wh-determiner
16.	PDT	Predeterminer	34.	WP	Wh-pronoun

17.	POS	Possessive ending	35.	WP\$	Possessive wh-pronoun
18.	PRP	Personal pronoun	36.	WRB	Wh-adverb

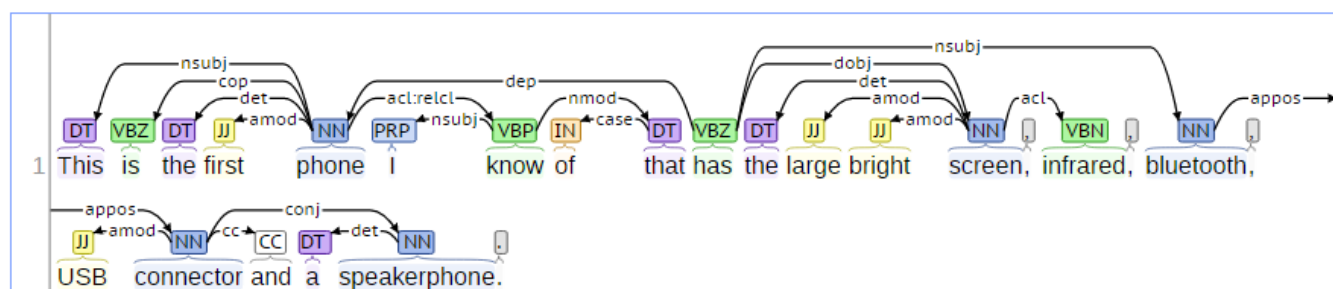
Table 8: POS Tags

Stanford POS Tagger

Stanford POS Tagger provides a simple representation of grammatical relationships in a sentence like Parts of Speech, Named Entity Recognition, Coreference, Basic Dependencies, Enhanced Dependencies.

Latest Stanford POS tagger is capable of providing 50 types of grammatical relationships like acomp: adjectival complement, amod: adjectival modifier, conj: conjunct, discourse: discourse

Basic Dependencies:



element etc.

Figure 32: Basic Dependencies

Enhanced Dependencies:

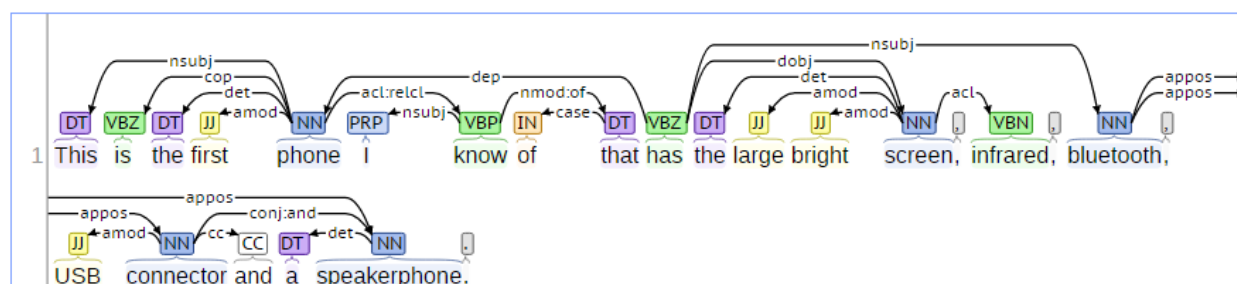
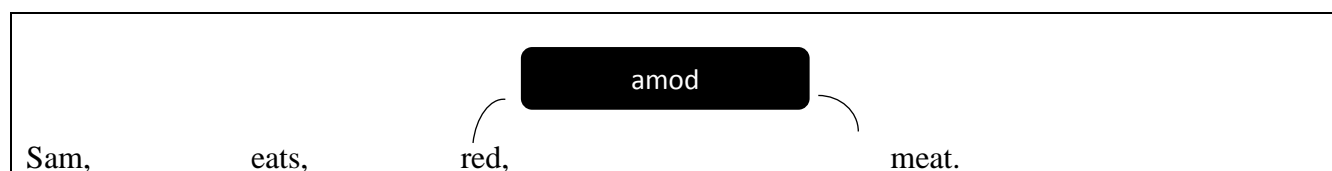


Figure 33: Enhanced Dependencies

Eg:-amod: adjectival modifier

amod provide the adjective phrase which modify the meaning of noun phrase.

“Sam eats red meat”



Maxent Tagger

Maxent Tagger is an already trained tagger and in this project, Maxent tagger is used with english-bidirectional-distsim.tagger.

english-bidirectional-distsim.tagger

English bidirectional distsim tagger is trained on WSJ sections 0-18 using a bidirectional architecture and including word shape and distributional similarity features.

6.3.8. Aspect / Feature Extraction

Aspects are the important features in a review comment. Aspect can be a word or phrase and mostly noun and noun phrases in a sentence. Following are the steps carried out to extract aspects.

- Take each POS tagged sentence
- Extract NN, NNP and NNS tagged words (repeated for all sentences)

Tag	Description
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural

- Count frequency of extracted words using Apriori algorithm
- Take most frequent aspects according to the assigned threshold value
- Save aspects in a file

In this project, we are focusing on the frequent aspects and to find the frequent aspects Apriori Algorithm is used.

6.3.9. Apriori Algorithm

Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases.

SPMF

SPMF is an open source data mining library written in java. It consists implementation for 120 data mining algorithms. It has the implementation for Apriori algorithm, which is a best known algorithm for identifying frequent item sets.

Support is the indicator of how frequently the item is appears in the database.

Support can be in two forms.

1. **absolute support** - absolute number of transactions which contains A
2. **relative support** - relative number of transactions which contains A

Frequent item set mining is used to find all frequent item sets using minimum support count. Firstly, noun and noun phrases in review sentence are identified and stored in a text file. Then minimum support threshold used to find all frequent aspects in the given review sentence and stored in a text file.

Following are the next steps of the proposed system.

1. Identify the aspect Environment, food
 2. Identify aspect related opinion word. – nice, bad
 3. Detect aspect orientation – positive, negative
- E.g.: - The environment is nice but food is bad.

6.4. Processing

In this step, preprocessed comments are sent to opinion mining process.

6.4.1. Remove Objective Comments

Remove objective comments filters out the sentences, which have at least an aspect in the saved aspect table. Otherwise, it will be marked as objective sentences and removed from the process. This will avoid the processing overhead of objective sentences which has no opinion and select only subjective sentences for further steps.

6.4.2. Opinion Words Identification

Opinion words express the opinion of an aspect. The important part of an aspect level OM application is to identify opinion words correctly. Mostly adjectives are the opinion words of a sentence. However, sometimes verbs, adverb adjective combinations and adverb verb combinations can be seen as opinion words. Following are the steps to extract opinion words.

- Searching 5-gram forwards and backwards from the aspect position in a sentence based on POS tag information. (Has some drawbacks when a sentence contains multiple aspects.)
- Using syntactic dependency relations between words to extract opinion words and calculate score of them.
 - Extract opinion words.

In the project Stanford Dependency parser is used to identify opinion words and then based on the results score is calculated using SentiWordNet.

Stanford Dependency Parser

Stanford Dependency Parser is a program works on the grammatical structure of a sentence. It identifies the subject, object of a sentence, grammatical relationships of the sentence etc. and shows how exactly the words have joined together to form the sentence. The latest Stanford POS tagger is capable of providing the relationships of the sentence. Therefore, the parser in POS tagger library is used in the project.

There are several trained models in the parser and here Default model of Dependency parse is used to find relationships.

Refer Stanford POS Tagger section for detail description of grammatical relationships of a sentence.

6.4.2.1. Find adjectives

To find adjectives in the sentence following Stanford type dependencies are used.

- **nsubj: nominal subject**

A nominal subject is a noun phrase, which is the syntactic subject of a clause. The governor of this relation might not always be a verb: when the verb is a copular verb, the root of the clause is the complement of the copular verb, which can be an adjective or noun (Marie-Catherine de Marneffe and Christopher D. Manning, 2015).

E.g.: -

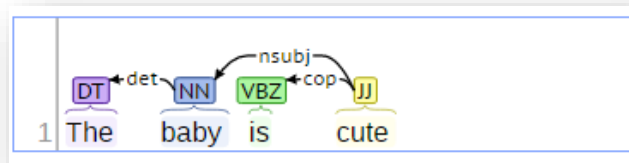


Figure 34: nsubj Dependency

- root (ROOT-0 , cute-4)
- det (baby-2 , The-1)
- nsubj (cute-4 , baby-2)
- cop (cute-4 , is-3)

- **amod: adjectival modifier**

An adjectival modifier is used to extract opinion words of the sentence. It is the relationship that shows any adjectival phrase that modifies the meaning of noun phrase (Marie-Catherine de Marneffe and Christopher D. Manning, 2015).

E.g.: -

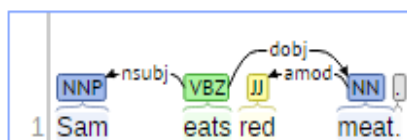


Figure 35: amod Dependency

- root (ROOT-0 , eats-2)
- nsubj (eats-2 , Sam-1)
- amod (meat-4 , red-3)
- dobj (eats-2 , meat-4)

- **advmod: adverb modifier**

An adverb modifier of a word is a (non-clausal) adverb or adverb-headed phrase that serves to modify the meaning of the word (Marie-Catherine de Marneffe and Christopher D. Manning, 2015).

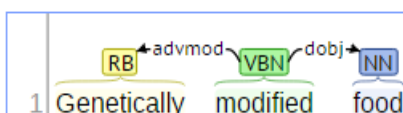


Figure 36: advmod Dependency

- root (ROOT-0 , modified-2)

- advmod (modified-2 , Genetically-1)
- dobj (modified-2 , food-3)

- **cc: coordination**

A coordination is the relation between an element of a conjunct and the coordinating conjunction word of the conjunct (Marie-Catherine de Marneffe and Christopher D. Manning, 2015).

- **conj: conjunct**

A conjunct is the relation between two elements connected by a coordinating conjunction, such as “and”, “or”, etc (Marie-Catherine de Marneffe and Christopher D. Manning, 2015).

- root (ROOT-0 , big-3)
- nsubj (big-3 , Bill-1)
- cop (big-3 , is-2)
- cc (big-3 , and-4)
- conj (big-3 , honest-5)

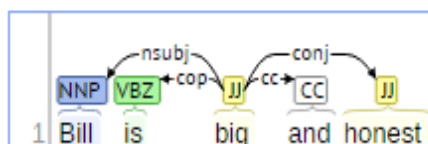


Figure 37: conj Dependency

- **root: root**

The root grammatical relation points to the root of the sentence. A fake node “ROOT” is used as the governor. The ROOT node is indexed with “0”, since the indexation of real words in the sentence starts at 1 (Marie-Catherine de Marneffe and Christopher D. Manning, 2015).

- root (ROOT-0 , man-5)
- nsubj (man-5 , Bill-1)
- cop (man-5 , is-2)
- det (man-5 , an-3)
- amod (man-5 , honest-4)

- **cop: copula**

A copula is the relation between the complement of a copular verb and the copular verb (Marie-Catherine de Marneffe and Christopher D. Manning, 2015).

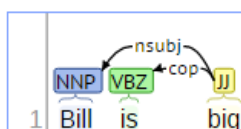


Figure 38: cop Dependency

- root (ROOT-0 , big-3)
- nsubj (big-3 , Bill-1)
- cop (big-3 , is-2)

- **acomp: adjectival complement**

An adjectival complement of a verb is an adjectival phrase, which functions as the complement (like an object of the verb) (Marie-Catherine de Marneffe and Christopher D. Manning, 2015).

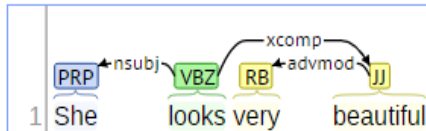


Figure 39: acomp Dependency

- root (ROOT-0 , looks-2)
- nsubj (looks-2 , She-1)
- advmod (beautiful-4 , very-3)
- xcomp (looks-2 , beautiful-4)

6.4.2.2. Find Adverbs

To find adverbs in the sentence following Stanford type dependencies are used.

- advmod : adverb modifier – described in the previous section
- advcl : adverbial clause modifier

An adverbial clause modifier of a VP or S is a clause modifying the verb (temporal clause, consequence, conditional clause, purpose clause, etc.) (Marie-Catherine de Marneffe and Christopher D. Manning, 2015).

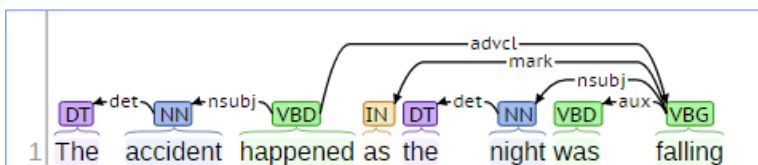


Figure 40: advcl Dependency

- root (ROOT-0 , happened-3)
- det (accident-2 , The-1)
- nsubj (happened-3 , accident-2)
- mark (falling-8 , as-4)
- det (night-6 , the-5)
- nsubj (falling-8 , night-6)
- aux (falling-8 , was-7)
- advcl (happened-3 , falling-8)

- amod : adjectival modifier – described in the previous section

Adjective adverb

Adverb does not have a clear meaning alone. But when they are paired with opinion word adjectives, it becomes an important factor when identifying the sentiment value of a sentence.

- Adverbs of affirmation – certainly, totally
- Adverbs of doubt – maybe, probably
- Strongly intensifying adverbs – exceedingly, immensely
- Weakly intensifying adverbs – barely, slightly
- Negation and minimizers - never

6.4.2.3.Find Verbs

To find verbs in the sentence following Stanford type dependencies are used.

- root : root
- nsubj : nominal subject
- dobj : direct object

The direct object of a VP is the noun phrase, which is the (accusative) object of the verb (Marie-Catherine de Marneffe and Christopher D. Manning, 2015).

- root (ROOT-0 , gave-2)
- nsubj (gave-2 , She-1)
- iobj (gave-2 , me-3)
- det (raise-5 , a-4)
- dobj (gave-2 , raise-5)

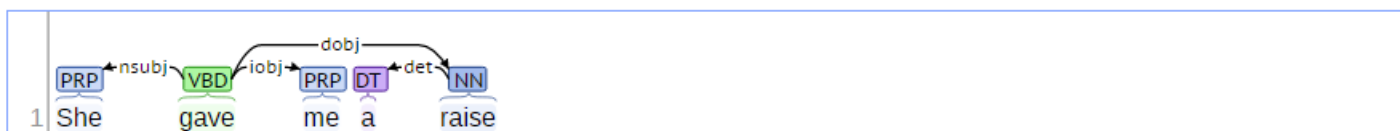


Figure 41: dobj Dependency

6.4.2.4.Find negation modifier

- neg : negation modifier

The negation modifier is the relation between a negation word and the word it modifies. (Marie-Catherine de Marneffe and Christopher D. Manning, 2015).

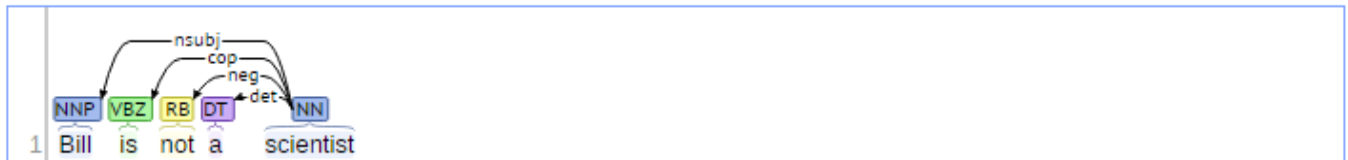


Figure 39: neg Dependency

- root (ROOT-0 , scientist-5)
- nsubj (scientist-5 , Bill-1)
- cop (scientist-5 , is-2)
- neg (scientist-5 , not-3)
- det (scientist-5 , a-4)

- pobj : object of a preposition – described in a previous section

Negation handling is apply negation rule if a word is near negation word,

- Negation negative → Positive
- Negation Positive → Negative

6.4.3. Calculating Scores

Score of opinion words are calculated in this step and to calculate scores weighed average is used.

Weighted Average

The weighted average is similar to an arithmetic mean of a set of numbers in which some elements of the set carry more importance (weight) than others.

$$\bar{x} = \frac{w_1x_1 + w_2x_2 + \cdots + w_nx_n}{w_1 + w_2 + \cdots + w_n}.$$

Figure 43: Weighted Average

Formulas used in score calculation

$$Initial\ Score = Initial\ Score + \frac{Synset\ Score(Positive\ Score - Negative\ Score)}{Sense\ Number}$$

$$Sum = Sum + 1 / Sense\ Number$$

$$Final\ Score = Initial\ Score / Sum$$

6.4.3.1.SentiWordNet

SentiWordNet is one of the most lexical databases for sentiment analysis. It contains opinion information on terms extracted from WordNet database by using a semi supervised learning method. This is publically available for research purpose. It contains words with its polarity scores - positive, negative, part of speech and glossary. It contains more than 117600 words.

SentiWordNet English is used by most of the researches and applications but it has several versions in several languages like, Italian etc.

Following is a brief description of SentiWordNet headers,

pos id	<ul style="list-style-type: none"> • POS - Part of speech tag like adjective, noun etc and id which is a unique id identifies the synset.
word	<ul style="list-style-type: none"> • the synset
positive score	<ul style="list-style-type: none"> • positive score assigned by the sentiwordnet to the synset
negative score	<ul style="list-style-type: none"> • negative score assigned by the sentiwordnet to the synset
sense no	<ul style="list-style-type: none"> • indicates how many senses are available for the particular synset
glossary	<ul style="list-style-type: none"> • example sentence using the synset

#	POS	ID	PosScore	NegScore	SynsetTerms	Gloss
a	00001740	0.125	0	able#1	(usually followed by 'to')	having the necessary means or skill or know-how or authority to do something; "able to do it"
a	00002098	0	0.75	unable#1	(usually followed by 'to')	not having the necessary means or skill or know-how; "unable to get to town without a car"
a	00002312	0	0	dorsal#2 abaxial#1	facing away from the axis of an organ or organism;	"the abaxial surface of a leaf is the underside or the back"
a	00002527	0	0	ventral#2 adaxial#1	nearest to or facing toward the axis of an organ or organism;	"the upper side of a leaf is known as the adaxial surface"
a	00002730	0	0	acrosopic#1	facing or on the side toward the apex	
a	00002843	0	0	basiscopic#1	facing or on the side toward the base	
a	00002956	0	0	abducting#1 abducent#1	especially of muscles; drawing away from the midline of the body or from an adjacent part	
a	00003131	0	0	adductive#1 adducting#1 adducent#1	especially of muscles; bringing together or drawing toward the midline of the body or toward the midline	
a	00003356	0	0	nascent#1	being born or beginning;	"the nascent chicks"; "a nascent insurgency"
a	00003553	0	0	emerging#2 emergent#2	coming into existence;	"an emergent republic"
a	00003700	0.25	0	dissilient#1	bursting open with force, as do some ripe seed vessels	
a	00003829	0.25	0	parturient#2	giving birth;	"a parturient heifer"
a	00003939	0	0	dying#1	in or associated with the process of passing from life or ceasing to be;	"a dying man"; "his dying wish"; "a dying wish"
a	00004171	0	0	moribund#2	being on the point of death; breathing your last;	"a moribund patient"
a	00004296	0	0	last#5	occurring at the time of death;	"his last words"; "the last rites"
a	00004413	0	0	abridged#1	(used of texts) shortened by condensing or rewriting;	"an abridged version"
a	00004615	0	0	shortened#4 cut#3	with parts removed;	"the drastically cut film"
a	00004723	0	0	half-length#2	abridged to half its original length	
a	00004817	0	0	potted#3	(British informal) summarized or abridged;	"a potted version of a novel"
a	00004980	0	0	unabridged#1	(used of texts) not shortened;	"an unabridged novel"
a	00005107	0.5	0	uncut#7 full-length#2	complete;	"the full-length play"
a	00005205	0.5	0	absolute#1	perfect or complete or pure;	"absolute loyalty"; "absolute silence"; "absolute truth"; "absolute alcohol"
a	00005473	0.75	0	direct#10	lacking compromising or mitigating elements; exact;	"the direct opposite"
a	00005599	0.5	0.5	unquestioning#2 implicit#2	being without doubt or reserve;	"implicit trust"
a	00005718	0.125	0	infinite#4	total and all-embracing;	"God's infinite wisdom"
a	00005839	0.5	0.125	living#3	(informal) absolute;	"she is a living doll"; "scared the living daylights out of them"; "beat the living hell out of him"

Figure 44: SentiWordNet

After finding opinion words and their weighted score using SentiWordNet aspect and sentence scores are calculated using below logics.

Adverb score is calculated using below logic.

```

1: function AGGREGATESCORE(AdverbScore
   AdverbScore , OpinionWordScore OpScore))
2:   AggregateScore : Variable which store the aggregate
   score of adverb and opinion word
3:   if OpScore = 0 then
4:     AggregateScore = 0
5:   else
6:     if AdverbScore > 0 then
7:       if OpScore > 0 then
8:         AggregateScore = min(1, OpScore + sf *
   AdverbScore)
9:       end if
10:      if OpScore < 0 then
11:        AggregateScore = min(1, OpScore - sf *
   AdverbScore)
12:      end if
13:    end if
14:    if AdverbScore < 0 then
15:      if OpScore > 0 then
16:        AggregateScore = max(-1, OpScore + sf *
   AdverbScore)
17:      end if
18:      if OpScore < 0 then
19:        AggregateScore = max(-1, OpScore - sf *
   AdverbScore)
20:      end if
21:    end if
22:  end if
23:  Return AggregateScore
24: end function

```

Figure 45: Aggregate Adverb Score Calculation Logic

Adjective and verb scores are calculated using below logic.

```

1: function   AGGREGATEADJECTIVESCORE(Adjectives,
   Ps)
2:   Adjectives : Array of adjectives
3:   AdjectiveScore : Priority Score of adjective retrieved
   from SentiWorNet.
4:   Adverbs : Array of Adverbs
5:   Agg_AdvAdj_Score : Aggregate score of adverb and
   adjective(Ex: 'very good').
6:   Agg_Adjective_Score : Aggregated scores of adjec-
   tives and Adverb Adjective combination.
7:   Agg_Adjective_Score = 0
8:   AdjectiveScore = 0
9:   if Adjectives  $\neq$  null then
10:    for all Adjective adj  $\in$  Adjectives do
11:      AdjectiveScore = Score(adj)
12:      if CheckNegation(adj, Ps)  $\neq$  False then
13:        Reverse the AdjectiveScore
14:      end if
15:      Adverb = getAdverbs(adj, Ps)
16:      if Adverb  $\neq$  null then
17:        AdverbScore = Score(Adverb)
18:        if CheckNegation(Adverb, Ps)  $\neq$  False
then
19:          Reverse the AdverbScore
20:        end if
21:        Agg_AdvAdj_Score +=
   AggregateScore(AdverbScore, AdjectiveScore)
22:      else
23:        AdjectiveScore + = AdjectiveScore
24:      end if
25:    end for
26:    if Adverb  $\neq$  null then
27:      Agg_AdvAdj_Score =
   Average(Agg_AdvAdj_Score)
28:      AdjectiveScore = Average(AdjectiveScore)
29:      Agg_Adjective_Score =
   Agg_AdvAdj_Score + sf * AdjectiveScore
30:    else
31:      Agg_Adjective_Score =
   Average(AdjectiveScore)
32:    end if
33:  else
34:    Agg_Adjective_Score = 0
35:  end if
36:  Return Agg_Adjective_Score
37: end function

```

Figure 46: Adjective/Verb Score Calculation Logic

6.5. Product Rating

After calculating positive and negative score, the lowest and highest scores for the sentence or aspect identified. Then the scores are divided in two equal parts and aspects, sentences are rated accordingly as below.

```

if(score >= 0){
    gap = (maxSentenceScore - 0)/5;
    if(score == 0){
        return 0;
    }
    if(score>= 0 &&score<= (gap)){
        return 1;
    }
    if(score>= (gap) &&score<= (2 * gap)){
        return 2;
    }
    if(score>= (2 * gap) &&score<= (3 * gap)){
        return 3;
    }
    if(score>= (3 * gap) &&score<= (4 * gap)){
        return 4;
    }
    if(score>= (4 * gap) &&score<= (maxSentenceScore)){
        return 5;
    }
}
else{
    gap = (minSentenceScore)/5;
    if(score< 0 &&score>= (gap)){
        return -1;
    }
    if(score< (gap) &&score>= (2 * gap)){
        return -2;
    }
    if(score< (2 * gap) &&score>= (3 * gap)){
        return -3;
    }
    if(score< (3 * gap) &&score>= (4 * gap)){
        return -4;
    }
    if(score< (4 * gap) &&score>= (maxSentenceScore)){
        return -5;
    }
}

```

Figure 47: Aspect / Sentence Rating Logic

6.6. UI Implementation

JSP, Java Script, CSS, HTML and images are used to design the UI. Following are the screens of System.

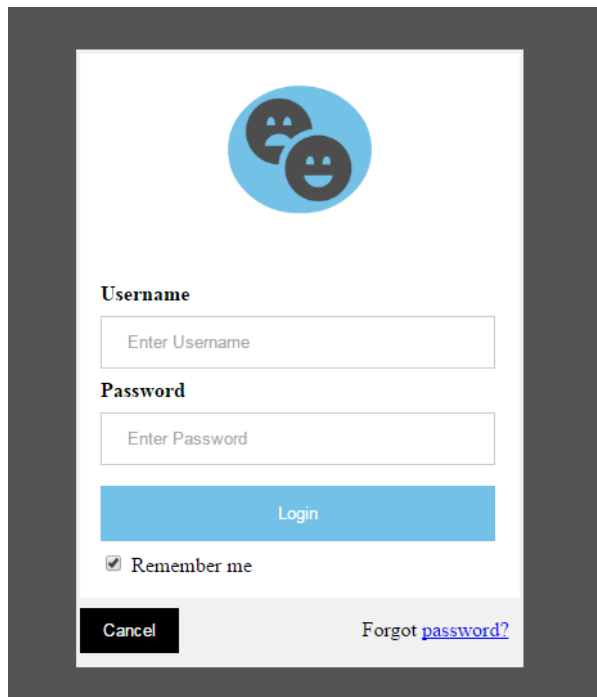
The login page features a central white card with a dark gray border. At the top of the card is a circular logo containing two stylized faces, one blue and one black. Below the logo, the form includes a 'Username' label, a text input field with the placeholder 'Enter Username', a 'Password' label, a text input field with the placeholder 'Enter Password', a blue 'Login' button, a checked 'Remember me' checkbox, a black 'Cancel' button, and a link for 'Forgot password?'.

Figure 48: Login Page

Once the user login the system, below Home page is prompted to the user.

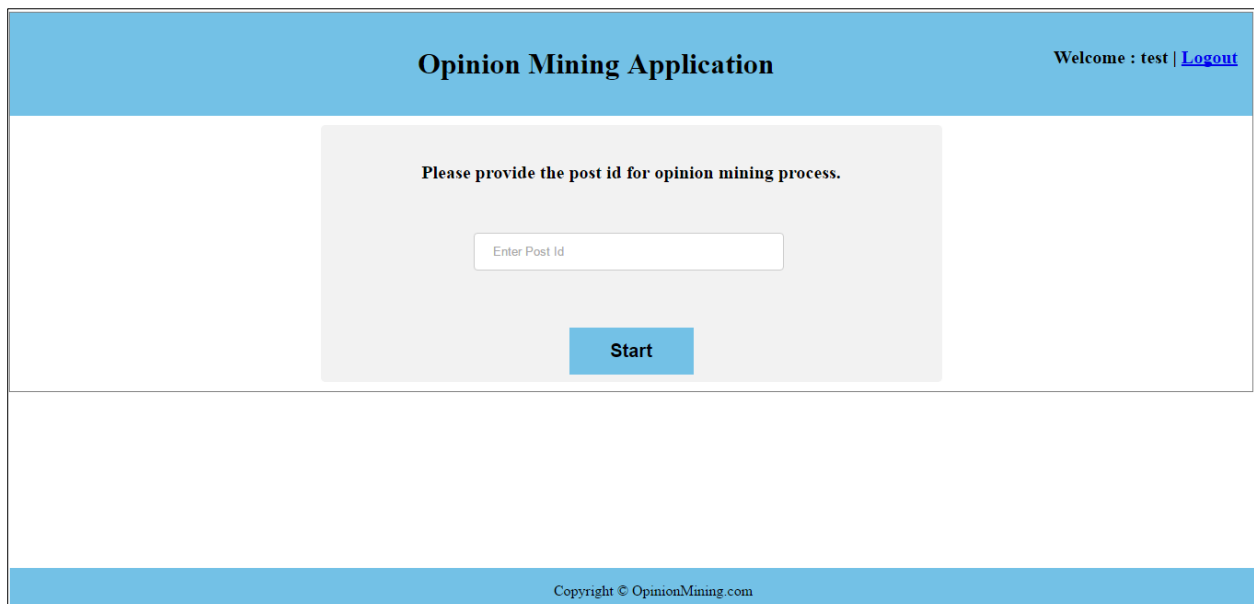
The home page has a blue header bar with the title 'Opinion Mining Application' and a user greeting 'Welcome : test | Logout'. The main content area is white and contains a gray box with the instruction 'Please provide the post id for opinion mining process.' Below this is a text input field with the placeholder 'Enter Post Id' and a blue 'Start' button. The footer is a blue bar with the text 'Copyright © OpinionMining.com'.

Figure 49: Home Page

Opinion Mining Results Page

[Back](#)

Opinion Mining Application

Welcome : test | [Logout](#)

Aspect Score Table

Aspect	Polarity	Final Score
disappointment	NEUTRAL	0.0
months	NEUTRAL	0.0
beauty	NEUTRAL	0.0
battery	NEUTRAL	0.0
life	POSITIVE	0.09000000000000001

Sentence Score Table

Sentence	Score	Polarity	Rating
only disappointment far has been battery life .	0.09000000000000001	POSITIVE	5.0
I 've had beauty nearly 2 months now I truely love it .	0.0	NEUTRAL	0.0

Copyright © OpinionMining.com

Figure 50: Aspect Result Table

Aspect	Polarity	Final Score
disappointment	NEUTRAL	0.0
months	NEUTRAL	0.0
beauty	NEUTRAL	0.0
battery	NEUTRAL	0.0
life	POSITIVE	0.09000000000000001

Sentence Score Table

Sentence	Score	Polarity	Rating
only disappointment far has been battery life .	0.09000000000000001	POSITIVE	5.0
I 've had beauty nearly 2 months now I truely love it .	0.0	NEUTRAL	0.0

Sentence Summary Table

No of Sentences	2
No of Positive Sentences	1
No of Negative Sentences	0
No of Neutral Sentences	1

Copyright © OpinionMining.com

Figure 51: Sentence Result Table and Summary

6.7. Deployment of the Application

The OM system is Spring Boot application written using spring framework. It is deployed and run in Pivotaltc Server locally. But it can deploy in any J2EE server. In the startup, it looks for an environment variable “APP_CONFIG” which points to the external configuration file. Therefore, before startup the server “APP_CONFIG” configuration value should be set as an environment variable or should pass as server argument. System can be access using the url - <http://localhost:8080/home> (localhost_url).

```
fb.comments.retreive.url=https://graph.facebook.com/v2.7/{post_id}/comments?
fb.comments.save.file.path=E:\\finalproject\\OMData\\appfiles\\fb-comments.txt

stop.word.file.path=E:\\finalproject\\OMData\\appconfig\\minimal-stop.txt
maxent.tagger.file.path=E:\\finalproject\\OMData\\appconfig\\english-bidirectional-distsim.tagger
sentiwordnet.file.path=E:\\finalproject\\OMData\\appconfig\\SentiWordNet_3.0.0_20130122.txt
aspect.file.path=E:\\finalproject\\OMData\\appfiles\\aspects.txt
aspect.save.file.path=E:\\finalproject\\OMData\\appfiles\\saved-aspects.txt

default.minium.support=0
default.weightage=0.3

fb.app.id=215183552230579
fb.app.secret=c752cc6e7c5ece362614a09b30353e00
fb.app.access.token=EAACEdEose0cBAOYOGGrR7XUA6jn6dZA2YP2qP5Fjg1OC8GDscTGjw9LoTsv7NO1riB5PmPE00aT2kmlk3ZBH5PV1bpTY03Y
```

Figure 52: Configuration File Content

6.8. User Manual

A user guide or user's guide, also commonly known as a manual, is a technical communication document intended to give assistance to people using a particular system. OM system is catered to different group of people like buyers, sellers, marking personal, marking researchers etc. Their computer literacy can be in different level. Therefore, user manual is prepared to assist them in using the system. It will also increase the user friendliness of the system. User manual is attached in Appendix section.

6.9. Summary

In this chapter a detail description has been given on all the tools, libraries, technologies and logics used to implement the system. Then deployment details and screen shots of the application are included. Finally, the user manual is attached for the reference.

7. Testing

Introduction

Testing Process

Testing Criteria

Testing

Issues Found

Limitation of Testing

Summary

7.1. Introduction

Testing is the process of evaluating the system is behaved as expected. As ANSI/IEEE 1059 standard, Testing can be defined as - A process of analyzing a software item to detect the differences between existing and required conditions (that is defects/errors/bugs) and to evaluate the features of the software item.

Usually testing is carried out by developers, testers, end users and it starts with the requirement analysis phase and continued until deploying the software.

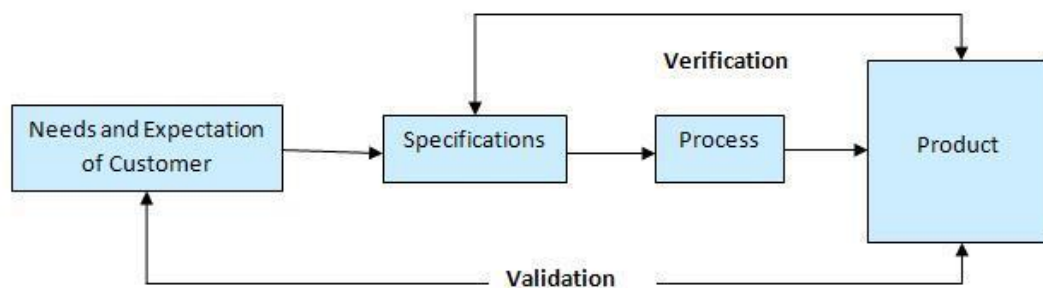


Figure 53: Validation and Verification in Testing

Testing can be divide to main two categories Verification and validation.

- Verification - "Are we building the product right?"
- Validation - "Are you building the right product?"

7.2. Testing Process

Testing process describes the flow followed throughout the testing process. Following are the activities carried out in Opinion Mining Application testing.

7.2.1. Planning and control

In this phase testing scopes and risks, testing approach and testing strategies are identified. In Opinion Mining application mainly functional testing – unit testing, integration testing and system testing are planned to carry out. In addition, nonfunctional testing, performance testing and usability testing will be carried out if there are necessary resources.

Functional Testing

Here it tests that Opinion Mining Application produced aspect results and sentence results as required.

Following testing are conducted to ensure the quality of OM application.

Unit Testing

In OM application, several JUnit unit test cases are written to test important units.

Integration Testing

In this application, several JUnit integration test cases are written to ensure its accuracy.

System Testing

OM application conducts the system testing using a manually tagged data set.

Nonfunctional testing

Nonfunctional testing is testing nonfunctional requirements like performance, usability, security, portability etc.

In opinion mining application, mainly performance and usability aspects are considered.

Performance testing

In our application the main performance testing is based on the speed that the third party libraries like, dependency parser and taggers are loaded in the system.

UI testing

In the application UI is tested by giving valid, invalid inputs and check whether appropriate error messages are popped to the user.

7.2.2. Analysis and design

Identify test conditions, design test cases, evaluate testability of the requirements , design , design test environment, identify infrastructure and tools is done in this phase.

Prepare a test data set is one of the most important step in this phase. For this app data set is retrieved from “www.cs.uic.edu” (Hu and Liu, KDD-2004). Most of the data in the data set is tagged as positive or negative and some of the data has the aspect it talks about. Data set is cleaned and preprocessed manually by removing unnecessary characters like “#”. Then the test data set is created by filtering positive and negative comments. (Filtered out the comparisons, indirect options, suggestions and improvements)

Following are the measures calculated according to the test results.

- Fraction of retrieved documents that are relevant to the query

$$\text{Precision} = \frac{\text{ExtractedValues} \cap \text{TrueValues}}{\text{ExtractedValues}}$$

- Fraction of documents that are successfully retrieved

$$\text{Recall} = \frac{\text{ExtractedValues} \cap \text{TrueValues}}{\text{TrueValues}}$$

$$\text{F measure} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

7.2.3. Implementation and Execution

In this phase, test cases and automation scripts are prepared. Then test cases are executed and expected and actual results are compared. Re testing is done if necessary.

Functional Test Cases

Test Case No	Test Case	Test Steps	Expected Result	Actual Result	Comments
1	Verify user is directed to Login Page when Home page is entered	Enter http://localhost:8080/home in the browser	User should be prompted the Login Page	Login page prompted	--
2	Verify user login functionality	Enter valid username and password Click login button	User should be prompted the home page Logged in user's username should be shown in the top right corner of the page	User successfully logged in. User's username is shown in the top right	--
3	Check invalid user login	Enter invalid username or password Click login button	User should be prompted an error message	Invalid Username and Password error message shown	--
4	Verify valid post id processing	Enter a valid post id Click on Process button	Processing image should be shown while processing the post comments. User should be shown the results page	Results are shown	--
5	Check empty post id	Click on Process button without entering a post id	User should be shown an error message "please	Error message is shown	

			enter a post id to process"		
6	Verify comments are turned to lower case	Check log	All comments should be in lower case	Comments are in lower case	Pre requisite TC - 4
7	Verify questions based are comments removed in pre processing	Check log	Question based comments should be removed	Questions are removed	Pre requisite TC - 6
8	Verify comments are tokenized	Check log	Comments should be tokenized	Comments are tokenized	Pre requisite TC - 7
9	Verify stop words are removed	Check log	Stop words should be removed	Stop words are removed	Pre requisite TC - 8
10	Verify comments are stemmed	Check log	Comments should be stemmed	Comments are stemmed	Pre requisite TC - 9
11	Verify comments are tagged	Check log	Comments should be POS tagged	Comments are tagged	Pre requisite TC - 10
12	Verify frequent item set is created	Set "default.minium.support" in config. Properties Check frequent items in saved-aspects.txt file	Frequent items should be in the file	Frequent items are in the file	Pre requisite TC - 11
13	Verify positive comments are identified as positive by the system	Upload positive comments Check the result	Comment should be identified as positive.	Check the results in 7.2.4 section	Pre requisite TC - 12
14	Verify negative comments are identified as negative by the system	Upload a negative comment Check the result	Comment should be identified as negative.	Check the results in 7.2.4 section	Pre requisite TC - 12
15	Verify aspects are identified by the system	Upload a comment which has a direct aspect Check the result	Aspect should be identified	Check the results in 7.2.4 section	Pre requisite TC - 12
16	Verify aspect polarity is correctly identified by the system	Check the result	Aspect polarity should be correctly identified	Check the results in 7.2.4 section	Pre requisite TC - 12
17	Verify logout	Click on logout button in the	User should be	User logout	Pre

	functionality	top right corner of the page	logout from the system User should be shown the login page with "logout success" message	from the system Successfully logout message shown	requisite TC - 2
18	Verify back button functionality	Click on the back button in top left corner of the page	User should be navigated to the previous visited page		Pre requisite TC - 2

Table 9: Functional Test Cases

Nonfunctional test cases

Test Case No	Test Case	Test Steps	Expected Result	Actual Result	Comments
1	Verify mandatory fields in Login page	Click on Login button without entering the username or password	Username is required, Password is required messages should be shown	Username is required message shown when username is not entered Password is required message shown when Password is not entered	
2	Verify mandatory fields in Home page	Click on Process button without entering the post id	Please enter the post id to continue the process message should be shown	Error message shown	
3	Verify dependency parser and tagger loading time	Check the dependency parser and tagger loading time during the server start up	Dependency parser and tagger should be loaded in less than 2 minutes time	Dependency parser and tagger loaded in less than 2 minutes	

Table 10: Non Functional Test Cases

7.2.4. Evaluating exit criteria and reporting

In this phase, test results are evaluated based on the measures like Precision, Recall and F measure that were described before.

Results on Aspect Identification

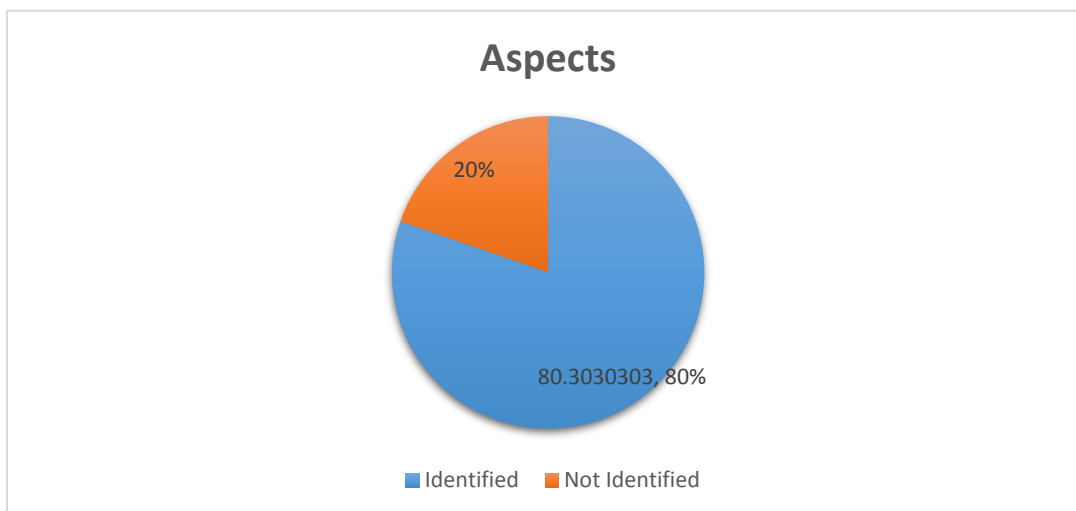


Figure 54: Aspect Identification Analysis

53 aspects are identified from 66 aspects in test data.

Results on Aspect Polarity Identification

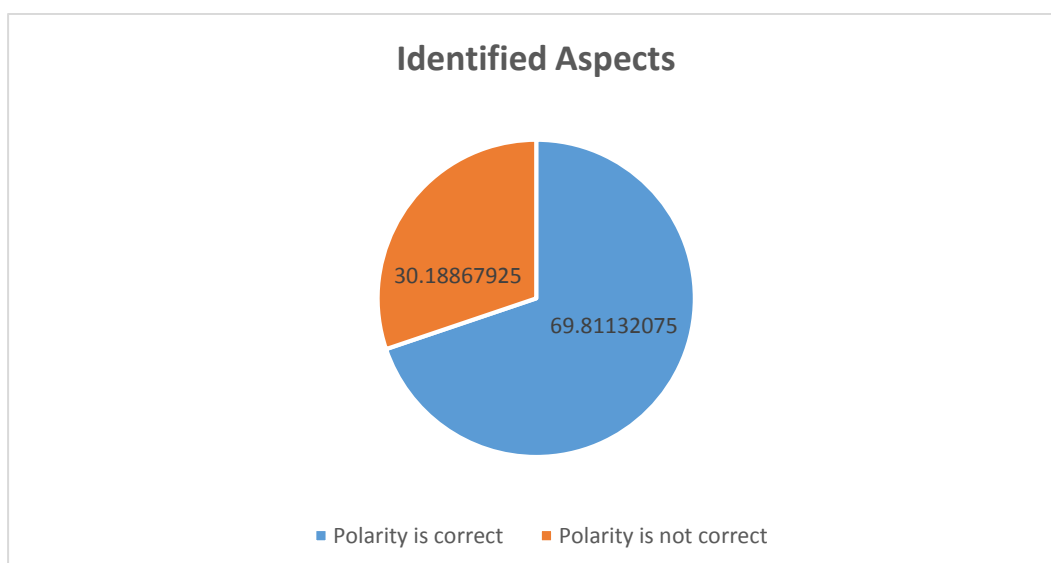


Figure 55: Aspect Polarity Identification Analysis

From identified 53 aspects 37 aspects are categorized as positive or negative correctly.

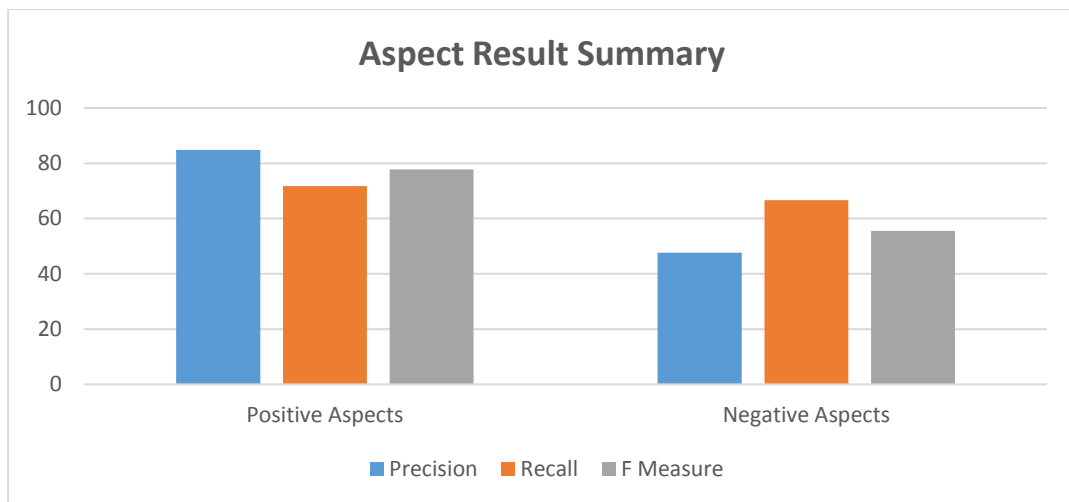


Figure 56: Aspect Result Summary Analysis

Results on Positive Comment Identification

Results on how positive and negative comments are identified through OM application.

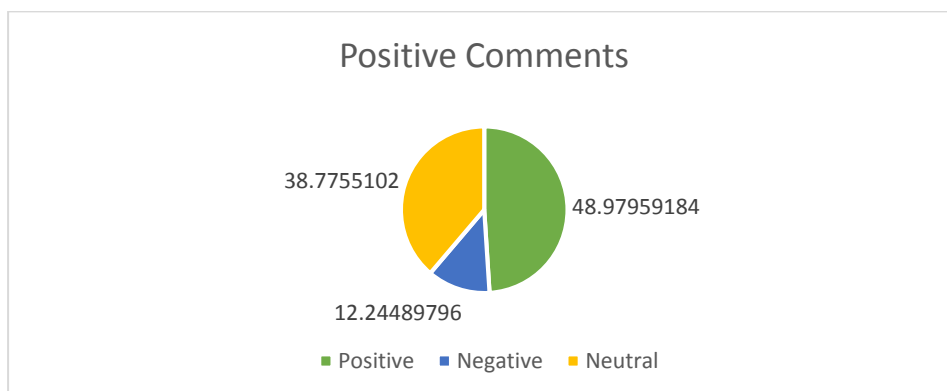


Figure 57: Positive Comments Analysis

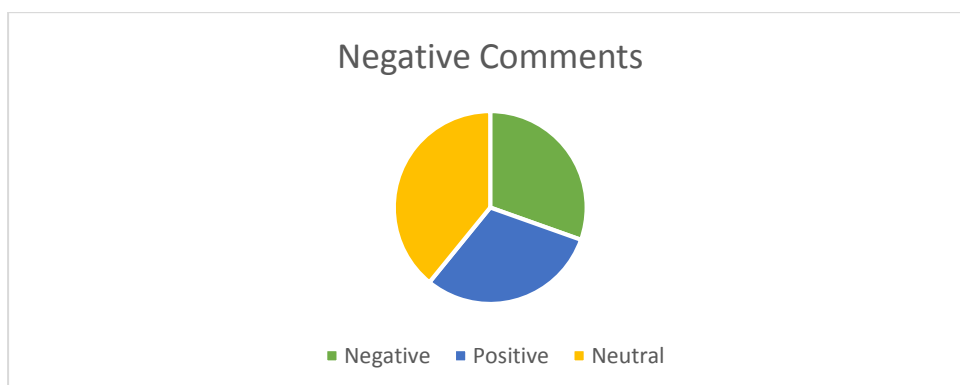


Figure 58: Negative Comments Analysis

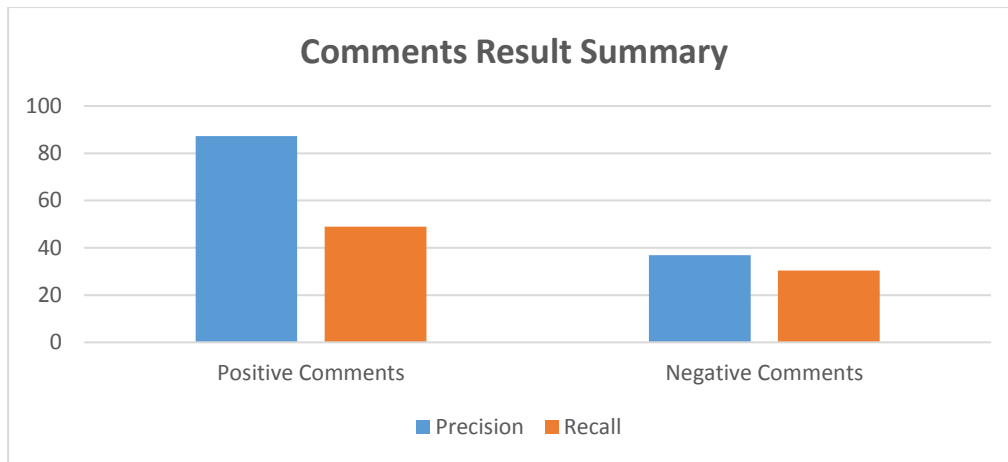


Figure 59: Comments Result Summary Analysis

7.2.5. Test Closure Activities

In OM application test closures is done after evaluating the results. Based on the results it is noted that aspect level opinion mining is in a satisfactory level and sentence level opinion mining should be improved. Furthermore, negative comments identification is in a lower level and a high attention should be given in that area.

In the future this application will be enhanced and tested until all the aspects identified correctly and all the direct opinion sentences are categorized correctly as POSITIVE and NEGATIVE.

7.3. Limitations of the testing process

All the testing is done in locally hosted application and therefore the performance values will be varied when deployed the application in real environment.

The test data is collected from “www.cs.uic.edu” (Hu and Liu, KDD-2004) and preprocessed before using it for testing. However, in the real environment there can be special characters in the comments, which cannot be processed in the application.

Following are some of the issues found while testing the application,

Issue	Solution
Some of the words are not in SentiWordNet	Assigned “0” as the value for not available words
Two word aspects are identified separately E.g.: - battery charger – as batter and charger	Need to address this issue in future

manual control – as manual and control	
Pictures, pics, which has the same meaning identified as different aspects	Need to address this issue in future
Some of the important words have been removed from stop word removal Eg: - I should give an A	Some of the words removed from stop word file but need to look in to this future , more deeply
Lot of domain specific words found in the test data	Need to maintain a domain specific word file for each domain in the future
Unnecessary aspects are found while identifying frequent aspects	User should be given a chance to select required aspects before sending them to opinion mining process
Reviews are written in an unstructured formant and there are spelling mistakes. Therefore, it does not get correct syntactic dependency.	Need to research on this area

Table 11: Issues and Solutions Table

7.4. Summary

Testing is the process of validating and verifying the software. The main functionalities of OM application are to identify the comments as POSITIVE or NEGATIVE and identify the aspects in the comments and identify them as POSITIVE and NEGATIVE. Functional testing is carried out to check whether comments are tagged accurately and whether aspects are found. Then UI and performance testing have been done to ensure its user friendliness and its speed. Nonfunctional testing result screen shots and test data used in the testing process are attached in appendix section.

8. Evaluation

Introduction

Evaluation Methodology

Selection of Evaluators

Evluation Findings

Self Evaluation

Summary

8.1. Introduction

Evaluation is the process of assessing the software and how its acceptance by the end users. In this chapter, the process used to evaluate the OM Application is described. The system is evaluated for accuracy and usability. Marketing personals are selected as the evaluators of the application and finally a self-evaluation has been done considering the overall project.

8.2. Evaluation Methodology

OM application is evaluated mainly using following methodologies.

- **Quantitative Assessment**

The accuracy of the results is assessed using quantitative measures like precision, recall and F measure. This is described in testing section and therefore here the focus is on qualitative assessment.

- **Qualitative Assessment**

Here the main focus is on evaluating the system using qualitative methods and system is evaluated using a questionnaire and prototype demonstration.

Refer appendix section for questionnaire used in the evaluation process and both open ended and close ended questions are included. Usability aspects are mainly evaluated using these questions.

Apart from the questionnaire prototype demonstration has been done prior to explain the functionalities of the system. It helped the evaluators to learn about the system, resolve ambiguities and discuss about the functionalities.

8.3. Selection of Evaluators

OM Application is designed for several group of people like prospective buyers of a product or service, sellers, marketing personal, business owners, people who are interested in particular product or service etc. Therefore, 10 people are selected as evaluators including all the categories mentioned above.

8.4. Evaluation Findings

Following facts are founded in evaluation.

- 70% like to recommend the system. It seems that this system is useful them to identify the aspects and their polarity – POSITIVE or NEGATIVE through the comments.

- 60% of the users satisfied with the results. Based on the testing and analysis results even though the aspect level opinion mining is satisfactory, sentence level opinion mining logic should be changed to increase the accuracy.
- Evaluators have suggested to summarize the sentence level results and proposed to give a figure on percentage of positivity or negativity on product or service by analyzing all the comments.
- Evaluators have suggested to have followings,
 - User should be able to save the results for later comparison
 - Application should be able to compare two products and give a comparison results
 - User should be able to select the features out of the system found aspects
- Based on the results the GUI is in a user acceptable level. However, there are some improvements to be done. By going through the screens, it is noticed that home page should be improved as it does not validate the post id patterns and therefore it leads to different issues.

8.5. Self-Evaluation

Self-evaluation is an important task as it helps to understand the gaps between expected and actual result. Self-evaluation is carried out considering the overall project and has evaluated whether functional, nonfunctional requirements are satisfied by the application. The project is developed using agile methodology and therefore the requirement analysis, designing, developing the system and testing phases are also evaluated in this phase. Following are the criterion's used in self-evaluation.

Criterion	Yes / No / Comment
Requirements are clearly identified and analyzed	Yes Firstly, by going through the current available opinion mining applications pros and cons identified. Then it is identified that current trend and focus is on aspect level opinion mining and therefore the researched is mostly focused on aspect level OM. In addition, it is identified that even though there are applications to analyze comments or review in twitter, there are no such to analyze Facebook post comments. Therefore, Facebook based OM application is planned to develop.
Application is properly designed	Need to improve the design Facebook application is integrated to OM application and currently it is accessed using user access tokens. However, user should be logged in to the Facebook application before login in to

	OM application. This flow should be designed to login the Facebook application after login the OM application and login should be handled internally via the OM application.
Application is implemented using industry best practices.	Yes Application is developed using MVC architecture and controllers, services are written as industry standards. All the property files are externalized and users are given the chance to configure the properties as required. Appropriate methods are written and code comments are written as possible. Latest technologies like spring boot, latest versions of Jackson libraries are used in the development.
Testing is adequate	No The main issue for testing is to find test data. The found data set does not directly used to test the application as it contains several unwanted characters, sarcasm based comments, suggestion etc. Therefore, the data set was preprocessed and filtered according to the format required. Then no of suitable test comments went down and it affected to testing process because larger data set is more appropriate than a smaller data set. In addition, the system has been tested using one data set. It would be better to test the system using different data sets as this system does not cater for a specific domain. (supports for all domains)
Usability of the system is satisfied	Usability of login and result pages are satisfied. Need to improve the Home page by validating post id, post id pattern etc.
Satisfied with the accuracy of the results	Accuracy of the aspect based opinion mining results is high and it can be considered as a satisfactory level. But the accuracy of sentence level aspect mining is bit low and need to improve the logic for better accuracy.

8.6. Summary

In this chapter implemented OM application is evaluated by public and myself. Qualitative evaluation is given more priority as previous chapter “testing” has done the quantitative analysis. Public is given a questioner and based on the results public opinion is evaluated. Finally, a self-evaluation is done to assess the developer point of view.

9. Conclusion

Introduction	
Discussion and Conclusion	
Contribution	
Learning Outcomes	
Recommendations and Future	
Concluding Remarks	
Summary	

9.1. Introduction

In this chapter, issues found in testing and evaluation chapters are discussed in detail. Then a conclusion is given after critically evaluating the results. Finally, suggestions and improvements are presented considering the whole system.

9.2. Discussion and Conclusion

Design and implement a Facebook based e commerce product rating was the main objective of this research. Product reviews or comments are a valuable source when purchasing a product or selling. However, because of the huge number of comments it has been difficult to find useful information and get a summarized idea.

By going through the current work, it has noticed that Facebook based opinion mining applications are less and therefore the decision has been made to design such system. Opinion mining can be done in three levels but the aspect level opinion mining is the current trend in opinion mining research area and it gives better results than other opinion mining levels sentence level, document level. Therefore, aspect based opinion mining study has been carried out to derive more knowledge from comments.

Aspect based opinion mining can be done using supervised learning methods, unsupervised methods etc. However, in this research lexicon based syntactic relation identification methodology is used to extract aspects. SentiWordNet 3.0 which is the enhanced version of SentiWordNet 1.0 is used as the lexicon. Frequent aspects are identified using Apriori algorithm and opinion words for the aspects are identified using syntactic relationships. Syntactic relationships are derived through Stanford Dependency Parser. Then negative and positive scores are calculated using SentiWordNet weighted average mechanism and finally the aspect polarity is identified along with the rating. Based on the aspects scores sentence score is also calculated and presented to the user in a summarized form. Positivity or negativity of a product can be varied according to user perception.

E.g.: - A user wants to buy a camera. If a user is more concerned on manual control of the camera and the manual control aspect is POSITIVE in the comments, it will be a positive product for him. Meantime if a user is more concerned on automatic settings and on automatic settings, aspect is NEGATIVE in the comments the same product will be a negative product for him. Therefore, product level summary is not given by doing a document level opinion mining. Since the aspects level summary is given user can decide whether the product is good or bad based on his requirements.

According to the testing results and evaluation findings even though the aspect polarity identification is in a satisfactory level, sentence level opinion mining should be improved. Sentence level score calculation logic should be reviewed. In addition, the negation handling should be improved as accuracy of negation identification is low. The process is tested using a single data set but it would be better to conduct more testing using different data sets.

While testing it is found that some of the important words are removed in preprocessing – stop word removal process. Stop words file should be revisited to avoid such situations in the future. Also when analyze the test data set it is identified that there are lot of domain specific words. It would be better to maintain domain specific word file with positive or negative score for each domain. Then it will improve the accuracy of scores for such words.

According to public evaluation, usability of the system is in a higher level and this application can be suggested to general public to use as an aid in decision making process. Because today everyone is busy, they do not have enough time go through all the positive, negative comments and make a comparison on them.

9.3. Contribution

There are lot of researches have been done on opinion mining as well as aspect level opinion mining. Syntactic approach using SentiWordNet has been researched and the system is developed integrating the Facebook platform. In this implementation, weighted average is used to calculate scores in SentiWordNet, which is an extended version of score calculation. In addition, lot of preprocessing steps are included in the application. It can be considered as an improvement in existing preprocessing flow.

9.4. Learning Outcomes

Theoretical knowledge gained from MSc programme have been practiced during the research and development in this project. The project created an opportunity to enhance analysis, critical evaluation skills and programming skills via the phases of this project like Literature review, system implementation etc.

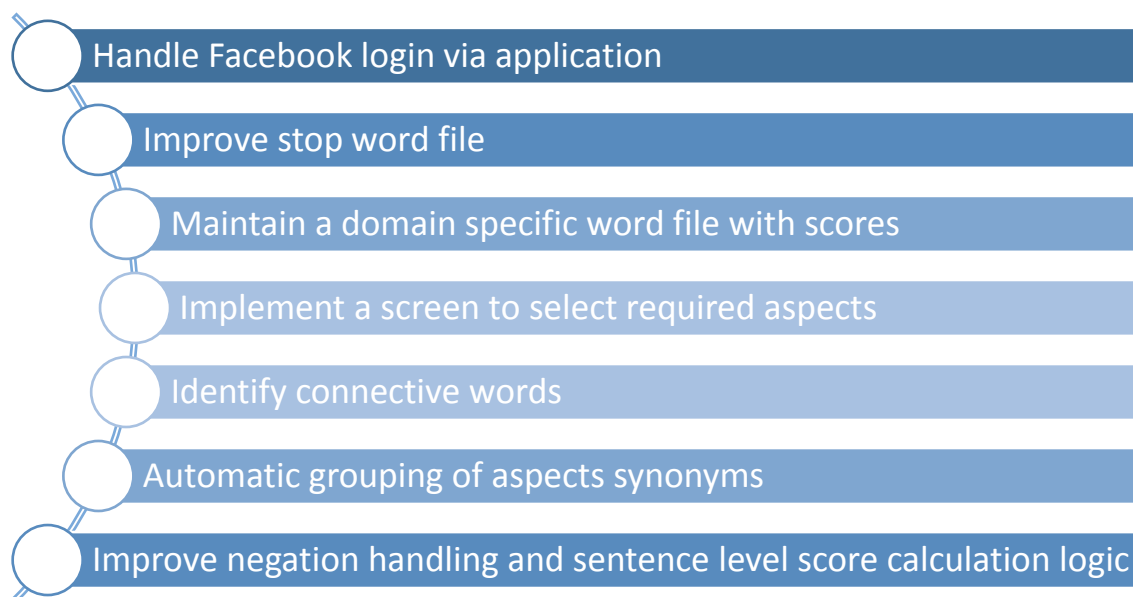
After selecting syntactic based approach using SentiWordNet, all the work has been challenging because opinion mining is a novel untouched area before. The first challenge was to design a process and then find the required libraries for POS tagging, dependency parsing, frequent item set mining

etc. Stanford libraries were used for the purpose. The next challenge was to integrate the Facebook platform for the system and spring social Facebook library is used to connect Facebook via Graph API using user access token. Valid user access token should be configured in the configuration file before using the system. Due to time constraints Facebook verification process (login process) is not handled inside the system and in the future it is planned be integrated. Then major challenge of this project was to create logics for opinion mining using grammatical relationships in the comments. Most of the logics are taken from the paper “A syntactic approach for aspect based opinion mining” (Chinsha T C and Shibily Joseph, 2015) and modified them when necessary.

In the testing phase by using precision, recall etc. To evaluate results helped to refresh the theoretical knowledge by practicing it in the project. Last but not least, by preparing the final documentation helped to improve the analysis and presentation skills immensely.

9.5. Recommendation and Future Work

Following are the recommendations identified after evaluating the Opinion Mining Application.



- **Handle Facebook login via application**

Currently Facebook data is accessed through user access token. To get the token user should login Facebook application and retrieved user access token should be saved in configuration file. But in the future Facebook login process will be handled inside the application. Once the user logged in to Opinion Mining application user should prompt a Facebook login page. Once user logged in, the

access token should be saved automatically and it should be used to authenticate next requests to Facebook. Then user access token should not be needed to save in the configuration file.

- **Improve stop word file**

During the testing, it is identified that some of the important words are removed in stop word removal process as stop word file contains such words.

E.g.: - I should give an A

Here “A” is not a stop word, but it is a word that shows the positivity. Therefore, need to analyze the stop word file furthermore and improve it to avoid such situations.

- **Maintain a domain specific word file with scores**

When analyzed the test data set it is identified that there are lot of domain specific words.

E.g.: - white balance, macros results etc.

Imagine a comment, “Low white balance.” Here we are not sure whether this is positive or negative. Therefore, better to keep a domain specific file with scores. When such aspect is found compare the score retrieved from SentiWordNet and created domain specific file for accuracy.

- **Implement a screen to select required aspects**

While testing it is found that unnecessary aspects are found in frequent item set mining. Therefore, user should be given a chance to select the aspects from system identified frequent aspect list. It will improve the efficiency of the process as it does not care about the unnecessary aspects but only consider about the user interested ones. In addition, it will increase the readability of the summary. New screen should be introduced to select required aspects like below.

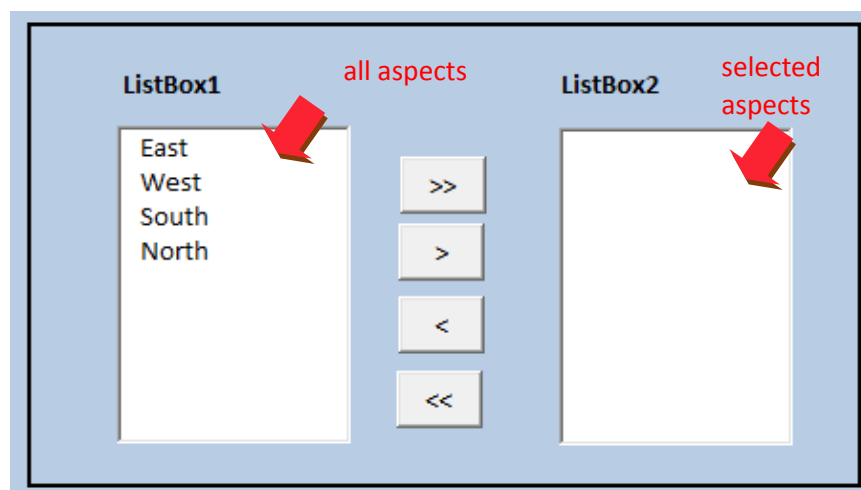


Figure 59: Aspects List Box

- **Identify connective words**

Identify connective words like AND, BUT, EITHER etc. during OM and make rules to used when assigning scores.

- **Automatic grouping of aspect synonyms**

In the test results, “pictures” and “pics” which has the same meaning identified as different aspects. Need to research on this area and need to design a logic to group the synonyms.

- **Improve negation handling and sentence level score calculation Logic**

During the test results evaluation, it is identified that negation handling and sentence score calculation logics should be improved. Need to perform more research on this area.

9.6. Concluding Remarks

Opinion mining is a popular research topic in recent past and most people are interested about the results of them. Due to highly unstructured nature of comments, it has been difficult to find a solution with 100% accuracy. Therefore, there is still a room for improvement and this research is also contributed in the area to identify the issues with some improvements in the process and results.

9.7. Summary

In this chapter conclusion of the project is discussed in detail. Then learning outcomes are described with the challenges faced. Finally, enhancements for the project are described as suggestions along with the possible solutions.

References

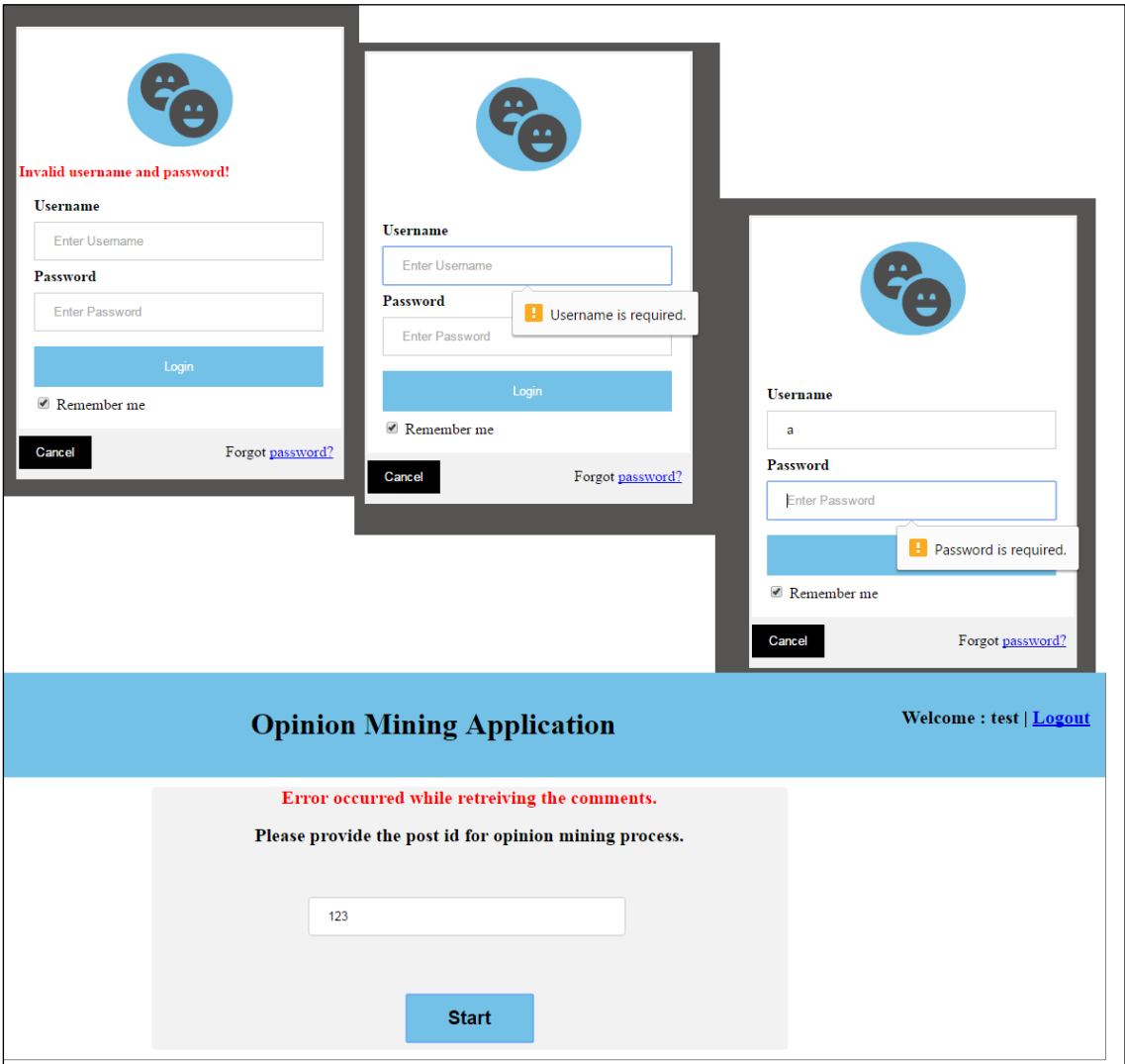
- Li, Li, Hwang, Yao, Zhang, Feifei, Guoliang Li, Seung-won, Bin, Zhenjie, 2014. Sarcasm Detection in Social Media Based on Imbalanced Classification. *Web-Age Information Management*, [Online]. Volume 8485 2014, 1. Available at: http://link.springer.com/chapter/10.1007/978-3-319-08010-9_49#page-1 [Accessed 24 November 2016].
- Wikipedia. 2016. *Social media*. [ONLINE] Available at: https://en.wikipedia.org/wiki/Social_media. [Accessed 24 November 2016].
- B. Raut; Londhe, Vijay, D. D., 2014. Opinion Mining and Summarization of Hotel Reviews. *2014 International Conference on Computational Intelligence and Communication Networks*, [Online]. DOI 10.1109/126 557 DOI 10.1109/CICN.2014.126, 556 - 559,. Available at: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7065546> [Accessed 22 January 2016].
- Jeyapriya, Selvi, A., C.S.Kanimozhi, 2015. Extracting Aspects and Mining Opinions in Product Reviews using Supervised Learning Algorithm. *IEEE SPONSORED 2ND INTERNATIONAL CONFERENCE ON ELECTRONICS AND COMMUNICATION SYSTEMS (ICECS '2015)*, [Online]. DOI: 10.1109/ECS.2015.7124967, 548 - 552. Available at: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7124967> [Accessed 26 January 2016].
- Ding; Le; Zhou; Wang; Shu, Juling; Zhongjian; Ping; Gensheng; Wei, 2009. An Opinion-Tree Based Flexible Opinion Mining Model. *2009 International Conference on Web Information Systems and Mining*, [Online]. DOI: 10.1109/WISM.2009.38, 149 - 152. Available at: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5369435> [Accessed 27 January 2016].
- Arora; Srinivasa, Rajeev; Srinath, 2014. A faceted characterization of the opinion mining landscape. *2014 Sixth International Conference on Communication Systems and Networks (COMSNETS)*, [Online]. DOI: 10.1109/COMSNETS.2014.6734936, 1 - 6. Available at: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6734936> [Accessed 31 January 2016].
- T C; Joseph, Chinsha; Shibily, 2015. A syntactic approach for aspect based opinion mining. *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*, [Online]. DOI: 10.1109/ICOSC.2015.7050774, 24 - 31. Available at: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7050774> [Accessed 31 January 2016].
- Cernian; Sgarciu; Martin, Alexandra; Valentin; Bogdan, 2015. Sentiment analysis from product reviews using SentiWordNet as lexical resource. *2015 7th International*

- Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, [Online]. DOI: 10.1109/ECAI.2015.7301224, WE-15 - WE-18. Available at: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7301224> [Accessed 2 February 2016].
- Liu;Lv;Wang, Lizhen; Zhixin; Hanshi , 2012. Opinion mining based on feature-level. *2012 5th International Congress on Image and Signal Processing*, [Online]. DOI: 10.1109/CISP.2012.6469929, 1596 - 1600. Available at: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6469929> [Accessed 4 February 2016].
 - Hamouda ;Rohaim, Alaa; Mohamed , 2012. Reviews Classification Using SentiWordNet Lexicon. *The Online Journal on Computer Science and Information Technology (OJCSIT)*, [Online]. Vol. (2) – No. (1), 120 - 123. Available at: <http://infomesr.org/attachments/123.pdf> [Accessed 5 February 2016].
 - Kumar;Sree Reddy, B.Sampath ;Dr.D.Bhanu , 2016. An Analysis on Opinion Mining: Techniques and Tools. *An Analysis on Opinion Mining: Techniques and Tools*, [Online]. Volume : 5 | Issue : 8, 489-492. Available at: <https://www.worldwidejournals.com/paripex/file.php?val=August 2016 1471848 868 163.pdf> [Accessed 30 August 2016].
 - Ghosh;Kar, Monalisa ;Animesh , 2013. Unsupervised Linguistic Approach for Sentiment Classification from Online Reviews Using Sentiwordnet 3.0. *International Journal of Engineering Research & Technology (IJERT)*, [Online]. Vol. 2 Issue 9, 55 - 60. Availableat: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.673.3049&rep=rep1&type=pdf> [Accessed 8 February 2016].
 - Catherine de Marneffe, D. Manning, Marie , Christopher, 2015. Stanford typed dependencies manual. *Stanford typed dependencies manual*, [Online]. Revised for the Stanford Parser v. 3.5.2 in April 2015, 1-28. Availableat: http://nlp.stanford.edu/software/dependencies_manual.pdf [Accessed 9 September 2016].
 - deMarneffe, D. Manning, Marie-Catherine, Christopher, 2008. Stanford typed dependencies manual. *Stanford typed dependencies manual*, [Online]. Revised for the Stanford Parser v. 3.5.2 in April 2015, 1-11. Available at: <http://nlp.stanford.edu/software/stanford-dependencies.shtml> [Accessed 3 October 2016].
 - easycalculation.com. 2016. *What is weighted average - Definition and Meaning*. [ONLINE] Available at: https://www.easycalculation.com/maths-dictionary/weighted_average.html. [Accessed 20 October 2016].
 - ISTQB EXAM CERTIFICATION. 2016. *What is fundamental test process in software testing?*. [ONLINE] Available at: <http://istqbexamcertification.com/what-is-fundamental-test-process-in-software-testing/>. [Accessed 22 November 2016].

- M. Lovelin, R., PonnFelciah, Anbuselvi, 2015. A study on sentiment analysis of social media reviews. *2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIECS)*, [Online]. 10.1109/ICIECS.2015.7192949, 1-3. Available at: <http://ieeexplore.ieee.org/document/7192949/> [Accessed 30 November 2016].
- Samha, Amani K, 2016. Aspect-Based Opinion Mining Using Dependency Relations. *International Journal of Computer Science Trends and Technology (IJCST)*, [Online]. Volume 4 Issue 1, Jan - Feb 2016, 113 - 123. Available at: <http://www.ijcstjournal.org/volume-4/issue-1/IJCST-V4I1P21.pdf> [Accessed 5 December 2016].

Appendices

Appendix A: GUI Validation Test Figures

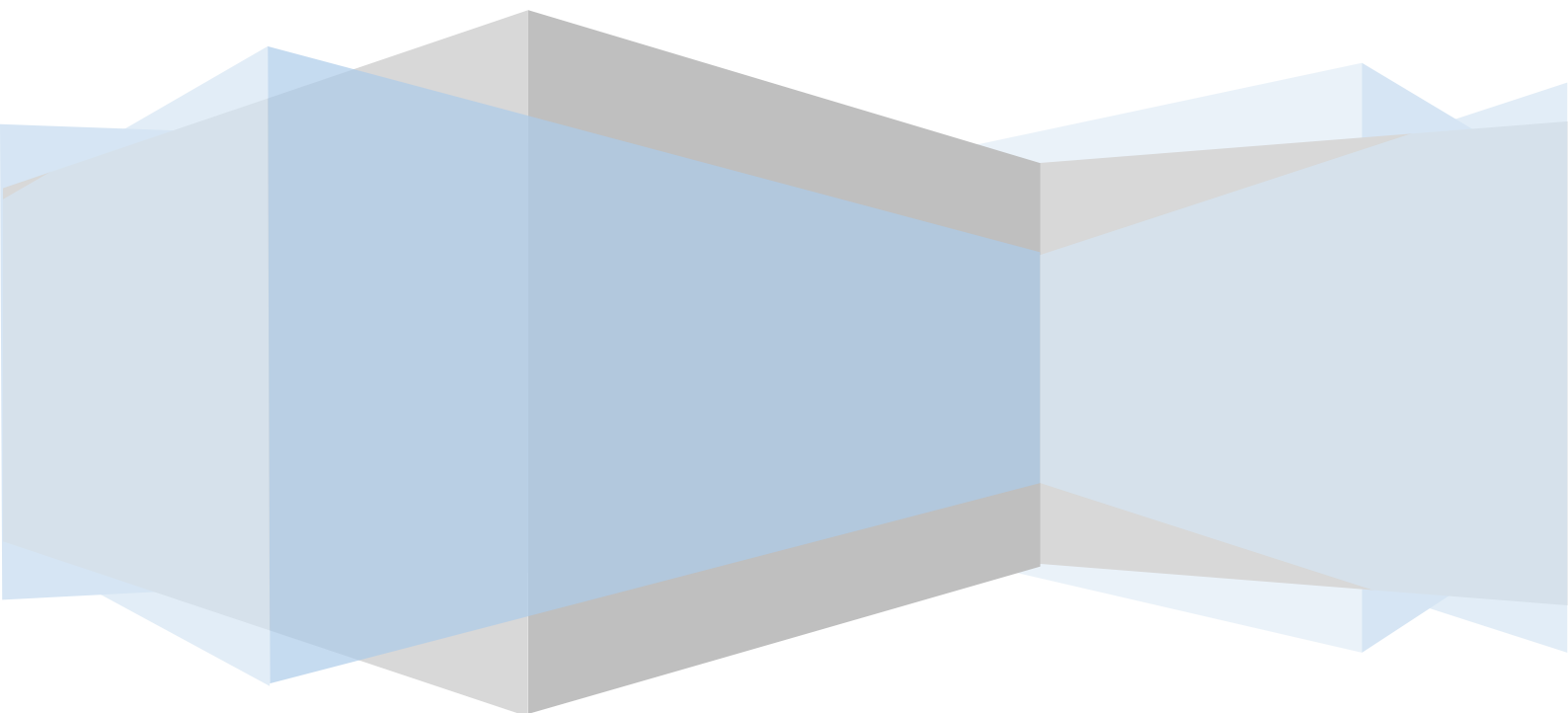


Appendix B: User Manual

Opinion Mining Application

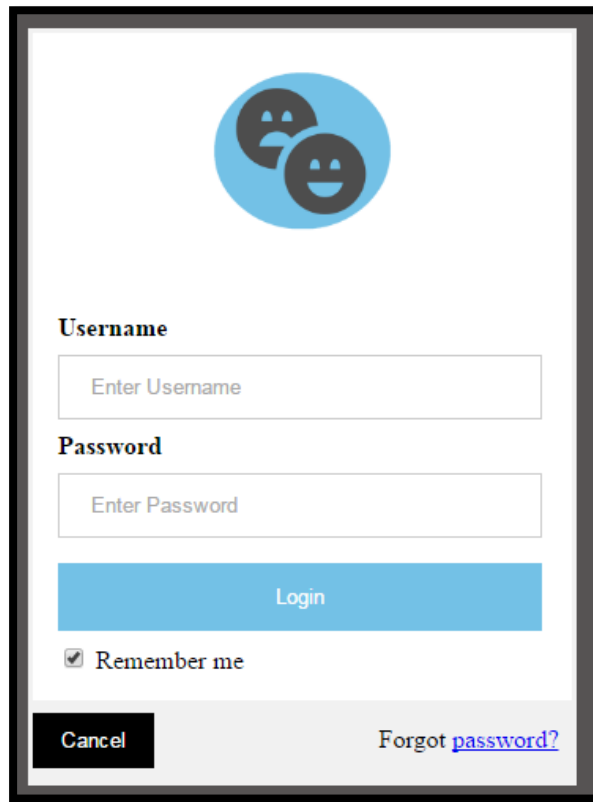
User Manual

Maneendra Madushani Perera



Steps to View Opinion Mining Results

1. Enter username, password and click Login button.



The login form features a blue circular icon with two stylized faces at the top. Below it, the label "Username" is followed by a text input field with the placeholder "Enter Username". The label "Password" is followed by a text input field with the placeholder "Enter Password". A blue "Login" button is positioned below the password field. A checkbox labeled "Remember me" is located below the login button. At the bottom left is a black "Cancel" button, and at the bottom right is a link that says "Forgot [password?](#)".

2. You will direct to Home page.
3. Enter the post id
Post id format should be → **userid_postid**, otherwise an error will thrown



The home page has a blue header with the text "Opinion Mining Application". Below the header, a light gray box contains the instruction "Please provide the post id for opinion mining process." followed by a text input field with the placeholder "Enter Post Id". A blue "Start" button is located at the bottom of this gray box.

4. Now you will see the results for the comments in entered post id

[Back](#)

Opinion Mining Application

Welcome : test | [Logout](#)

Aspect Score Table

Aspect	Polarity	Final Score
disappointment	NEUTRAL	0.0
months	NEUTRAL	0.0
beauty	NEUTRAL	0.0
battery	NEUTRAL	0.0
life	POSITIVE	0.09000000000000001

Sentence Score Table

Sentence	Score	Polarity	Rating
only disappointment far has been battery life .	0.09000000000000001	POSITIVE	5.0
I 've had beauty nearly 2 months now I truely love it .	0.0	NEUTRAL	0.0

Aspects Results

Copyright © OpinionMining.com

Aspect	Polarity	Final Score
disappointment	NEUTRAL	0.0
months	NEUTRAL	0.0
beauty	NEUTRAL	0.0
battery	NEUTRAL	0.0
life	POSITIVE	0.09000000000000001

Sentence Score Table

Sentence	Score	Polarity	Rating
only disappointment far has been battery life .	0.09000000000000001	POSITIVE	5.0
I 've had beauty nearly 2 months now I truely love it .	0.0	NEUTRAL	

Sentence Results

Sentence Summary Table

No of Sentences	2
No of Positive Sentences	1
No of Negative Sentences	0
No of Neutral Sentences	1

Copyright © OpinionMining.com

Appendix C: Gantt chart

	Task Name	Duration	Start Date	End Date	Q4			Q1			Q2			Q3		
					Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep
1	Initial Phase	59d	11/01/16	01/03/17	Initial Phase											
2	Topic selection	15d	11/01/16	11/15/16												
3	Background search on topic	15d	11/16/16	12/01/16												
4	Prepare draft Terms of Reference	13d	12/02/16	12/14/16												
5	Prepare final Terms of Reference	16d	12/15/16	01/01/17												
6	Literature Review	54d	01/04/17	02/28/17	Literature Review											
7	Review research papers	27d	01/04/17	01/31/17												
8	Critique the literature	14d	02/01/17	02/14/17												
9	Prepare literature review document	13d	02/15/17	02/28/17												
10	Requirement Analysis	17d	03/01/17	03/20/17	Requirement Analysis											
11	Analyse requirements	7d	03/01/17	03/08/17												
12	Draft Requirement Specification	7d	03/09/17	03/16/17												
13	Finalize Requirement Specification	3d	03/17/17	03/20/17												
14	Design and Development	74d	03/21/17	06/10/17	Design and Development											
15	Design the system	7d	03/21/17	03/30/17												
16	Develop Prototype	7d	03/31/17	04/07/17												
17	Submit Interim Report	9d	04/08/17	04/17/17												
18	Develop System	37d	04/18/17	05/25/17												
19	Write unit test cases	7d	05/26/17	06/03/17												
20	Developer testing	7d	06/03/17	06/10/17												
21	Testing	15d	06/11/17	06/30/17	Testing											
22	Develop test plan	6d	06/11/17	06/17/17												
23	Complete unit testing	2d	06/18/17	06/20/17												
24	Conduct integration testing	1d	06/21/17	06/22/17												
25	Conduct functional testing	2d	06/23/17	06/25/17												
26	Fix issues	4d	06/26/17	06/30/17												
27	Documentation	29d	07/01/17	08/01/17	Documentation											
28	Prepare draft project report	14d	07/01/17	07/14/17												
29	Get the feedback	1d	07/15/17	07/16/17												
30	Prepare final project report	14d	07/17/17	07/31/17												

Appendix D: Questionnaire Used for Evaluation

- How likely is it that you would recommend this software?

Not at all likely

extremely likely

1	2	3	4	5
---	---	---	---	---

- Do you satisfied with the results provided by the software?

☐ Strongly Agree

☐ Agree

☐ Neutral

☐ Disagree

☐ Strongly Disagree

- What are the areas that you would think an improvement is needed?

.....

- What are the additional features you would like to have in the software?

.....

- Do you satisfy with the GUI of the software?

☐ Strongly Agree

☐ Agree

☐ Neutral

☐ Strongly Disagree

☐ Disagree