

Final Year Project Report

# **SENTIMENTS ANALYSIS ON URDU TEXT**

**BCIT (Batch: 2010-11)**

## **Project Advisor**

Dr.Sohail Abdul Sattar

Co-Chairman  
CSIT Dept  
NEDUET

## **Submitted by**

Nimra Ahmad

CT-10004

Qunoot Ahmed

CT-10021

Rabail M.Ali

CT-10040

Nida Saeed

CT-10050



**DEPARTMENT OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY**  
**NED University of Engineering and Technology, Karachi.**

## **PREFACE**

---

Sentiments are the center of all human activities. Sentiment analysis or opinion mining is that technique that analyzes people's sentiments, opinions and attitudes from written text. In English you can find a lot of work done on sentiments analysis but in Urdu it has not given much attention till yet.

This report gives the detailed explanation of our project “Sentiments Analysis in Urdu Text”. It’s a research based project which is attempted by us. In this report we have given the introduction to Sentiments Analysis, basic tools and technology used in this project, the logic behind this project, system’s architecture, its sub modules, its states all have been discussed.

Important results and conclusion are derived by doing the Sentiment Analysis of our collected data. Future enhancements and recommendations are also given for the people who are interested to work upon it.

## **ACkNOWLEDGMENT**

---

First of all, we thank Almighty Allah who gives us the strength and ability to think, work and deliver what we are assigned to do. Secondly, we must be grateful to our internal supervisor Dr.Sohail Abdul Sattar who guided us in this project. We also acknowledge our teachers who guided, taught and helped us during our whole study period, departmental staff, university staff or other then this.

## INTRODUCTION TO GROUP MEMBERS

---



Nimra Ahmad CT- 10004  
(Specialization in C#, SQL)  
Immediate Contact: (0331-2125477, nimra\_ahmad@hotmail.com)



Qunoot Ahmed CT- 10021  
(Specialization in C#, SQL)  
Immediate Contact: (0333-3545346, qunoot\_ahmed@hotmail.com)

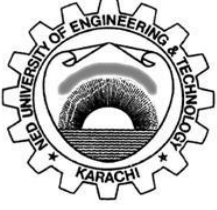


Rabail M. Ali CT- 10040  
(Specialization in C#, SQL)  
Immediate Contact: (0343-2846146, rabail\_40@yahoo.com)



Nida Saeed CT- 10050  
(Specialization in C#, SQL)  
Immediate Contact: (0336-2507299, nidasaeed1113@yahoo.com)

*NED University of Engineering and Technology, Karachi*



**DEPARTMENT OF COMPUTER SCIENCE &  
INFORMATION TECHNOLOGY**

**CERTIFICATE OF COMPLETION**

This is to certify that the following students

Nimra Ahmad	CT-10004
Qunoot Ahmed	CT-10021
Rabail M.Ali	CT-10040
Nida Saeed	CT-10050

have successfully completed their final year project titled

**SENTIMENTS ANALYSIS ON URDU TEXT**

In the partial fulfillment for the requirements of the Degree of Bachelor of Computer Science & Information Technology during the academic session 2010-2011.

---

**Dr.Sohail Abdul Sattar**  
Co-Chairman  
CSIT Dept  
NEDUET



*NED University of Engineering and Technology, Karachi*

**DEPARTMENT OF COMPUTER SCIENCE &  
INFORMATION TECHNOLOGY**

**CERTIFICATE OF COMPLETION**

This is to certify that the following students

Nimra Ahmad	CT-10004
Qunoot Ahmed	CT-10021
Rabail M.Ali	CT-10040
Nida Saeed	CT-10050

have successfully completed their final year project titled

**SENTIMENTS ANALYSIS ON URDU TEXT**

**IN THE PARTIAL FULFILLMENT FOR THE REQUIREMENTS OF THE DEGREE OF  
BACHELOR OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY DURING THE  
ACADEMIC SESSION 2010-2011**

**DR.NAJMI GHANI HAIDER**

**CHAIRMAN**

**CSIT DEPARTMENT**

**NED UNIVERSITY**

## SYNOPSIS

---

In this report we have ten chapters that describe our project Sentiment Analysis in Urdu Text in detail.

**Chapter 1** is the “Introduction”, in this chapter; we have given an overview of our project describing the inputs and outputs of the system, its domain, its market value. It also includes the history of the project and any prior work that is done on this system. It also consists of the tools and technology used to build the project.

**Chapter 2** is “User End”; this chapter will provide the details that how the users can interact with our system and how they can use this system to perform Sentiments Analysis on Urdu text. Intended users of the system are identified, Uses Cases diagrams for the system along with their description are given. It also constitutes of special considerations for using this system including data requirements and software requirements.

**Chapter 3** is “Functional and Data Description”; this chapter describes the architecture of the system. It contains the functionality of the system and type of information that flows through the system. It also consists of activity diagram of the system. Also contains information about the data objects and the sub modules of the system.

**Chapter 4** is “Sub systems/Modules”; in this chapter detailed explanation of each sub module is given along with the data flow diagram of each module.

**Chapter 5** is “Behavior Model and Description”, this chapter discusses behavior of the system. Different states of the system are explained and event diagrams are drawn.

**Chapter 6** is “System Prototype Modeling and Simulation Results, this chapter contains that the detail about the prototype you have designed before building the actual model. We have used the Waterfall model.

**Chapter 7** is “System Estimates and Actual Outcome”, this chapter

Discusses the estimations that we proposed earlier and the actual out comes after completion of project.

**Chapter 8** is “Testing”, this chapter contains the test plan, test cases you developed and testing methods applied.

**Chapter 9** is “Future Enhancements and Recommendations”; this chapter contains description about the deviation observed from the proposed system that can be recommended as future work.

**Chapter 10** gives the Conclusion and the Results.



# CONTENTS

---

## PREFACE

- A. ACKNOWLEDGEMENTS
- B. INTRODUCTION TO GROUP MEMBERS
- C. CERTIFICATE OF COMPLETION BY PROJECT ADVISOR
- D. CERTIFICATE OF COMPLETION BY CHAIRMAN OF DEPARTMENT
- E. CERTIFICATE OF COMPLETION BY EXTERNAL ORGANIZATION (IF EXTERNAL PROJECT)
- F. SYNOPSIS
- G. TABLE OF CONTENTS
- H. LIST OF FIGURES
- I. LIST OF TABLES

## Chapter 1 Introduction

1.1	What is Sentiments Analysis?	16
1.2	Goals and Objectives	17
	1.2.1 Applications in Business	18
	1.2.2 Applications in Business Intelligence	18
	1.2.3 Political Sentiment Analysis	18
1.3	System statement of scope	19
	1.3.1 System description	19
	1.3.2 Our approach	19
	1.3.3 Major inputs/processing functionality and outputs	20
	1.3.4 Data collection	20
	1.3.5 Data dictionary	20
1.4	System context	21
	1.4.1 Use of the system in business	21
	1.4.1.1 Business questions and Sentiment analysis answers	22
	1.4.1.2 How can it be used for business intelligence	22
1.4.2	Prospects of the system in market	22
	1.4.2.1 Social marketing	23

1.4.3	Use by common people	24
1.4.4	Strategic issues	24
1.5	Theoretical background	25
1.5.1	Background material provided	26
1.6	Technology and Tools	27
1.6.1	Technology	27
1.6.1.1	“Bag of words” model	27
1.6.1.2	Natural Language Processing, and the attempt to truly “understand” the text	27
1.6.1.3	What is Artificial Intelligence?	27
1.6.1.4	What is Machine Learning?	28
1.6.1.5	What is Natural Language Processing (NLP)?	28
1.6.1.6	Rule-based approach	28
1.6.2	Tools	29
1.6.2.1	Weka	29
1.6.2.2	Rapid Miner	30
1.6.2.3	Visual Studio Windows Forms	30
1.6.2.4	MS Access	31
<b>Chapter 2</b>	<b>Usage Scenario/User Interaction</b>	<b>32</b>
2.1	User profiles	32
2.2	Intended users	33
2.3	Special usage considerations	33
2.3.1	Data requirements	33
2.3.2	Software requirements	33
2.4	Use cases	34
2.4.1	How overall system works?	34
2.4.2	Basic concept of the system	35

2.4.3	Frequency of Positive and Negative words	36
-------	--	----

### **Chapter 3      Functional and Data description**

#### **39**

3.1	System architecture	39
3.1.1	Architecture model	39
3.1.2	Subsystem/modules overview	43
3.2	Data description	44
3.2.1	Major Data objects	45
3.2.1.1	Sentiment lexicon generation	45
3.2.1.2	Different types of sentences	45
3.3	Major Data objects	48
3.3.1	System Interface description	48

### **Chapter 4      Subsystems/Modules**

#### **53**

4.1	Description for Data Input Module	53
4.1.1	Subsystem scope	53
4.1.2	Subsystem flow diagram	53
4.2	Description for Database container module	54
4.2.1	Subsystem scope	54
4.3	Description for Result module	55
4.3.1	Subsystem scope	55
4.4	Description for Frequency module	56
4.4.1	Subsystem scope	56
4.5	Description for Graphical module	57
4.5.1	Subsystem scope	57
4.6	Description for Data separator module	58
4.6.1	Subsystem scope	58

### **Chapter 5      Behavior Model and Description**

#### **59**

5.1	Description for System behavior	59
-----	---------------------------------	----

	5.1.1	Events/ Interrupts	59
	5.1.2	States	60
	5.2	State transition diagram	61
<b>Chapter 6</b>		<b>System Prototype Modeling and Simulation result</b>	<b>63</b>
	6.1	History of the Waterfall Model	63
	6.2	What is Waterfall Model?	63
	6.3	When should we use the Waterfall Model?	63
	6.4	Features of the Waterfall Model	64
	6.5	How Waterfall Model used in our project?	65
	6.5.1	Requirement gathering and Analysis	65
	6.5.2	System design	66
	6.5.3	Implementation	67
	6.5.4	Testing	69
	6.5.5	Deployment of System	69
	6.5.6	Maintenance	69
	6.6	Disadvantages of the Waterfall Model	70

	6.7	Waterfall Model Application	70
<b>Chapter 7</b>		<b>System Estimates and Actual Outcome</b>	<b>72</b>
	7.1	Historical Data used for estimates	72
	7.2	Estimation techniques applied and results	72
	7.2.1	Breaking down	72
	7.2.2	Similar value	73
	7.3	Actual Results and Deviations from Estimates	73
	7.4	System resources (required and used)	74
	7.4.1	System resources required	74
	7.4.2	System resources used	75
<b>Chapter 8</b>		<b>Testing</b>	<b>76</b>
	8.1	System test and procedure	76
	8.2	Testing strategy	78
	8.2.1	Unit Testing	78
		8.2.1.1 Unit testing procedure	79
		8.2.1.2 Testing each Module	79
		8.2.1.3 Unit cases of our	79
		system	79

8.2.2	Integration Testing	
80		
8.2.2.1	Order of Integration by System function	
81		
8.2.2.2	Integration test conditions	
81		
8.2.3	Validation testing	
82		
8.2.4	High-order testing (a.k.a. System Testing)	
82		
8.2.4.1	Security testing	
82		
8.2.4.2	Stress testing	
83		
8.3	Testing resources and staffing	
83		
8.4	Test metrics	
83		
8.5	Testing Tools and environment	
83		
8.6	Test record keeping and test log	
84		
<b>Chapter 9</b>	<b>Future Enhancements and Recommendations</b>	
	<b>85</b>	
9.1	Proposed Future work	85
9.2	Aspect-Based Sentiments Analysis	87
<b>Chapter 10</b>	<b>Conclusion</b>	
	<b>89</b>	
<b>APPENDICES</b>		<b>91</b>
<b>REFERENCE</b>		<b>94</b>
<b>GLOSSARY</b>		<b>95</b>

## LIST OF FIGURES

CHAPTER NO.	FIGURE NO.	Name of the Figure	PAGE NO.
2	2.1	Use Case Diagram for the working of the whole system	34
	2.2	Use Case Diagram for the basic working of the system	35
	2.3	Use Case Diagram for Frequency Checking	36
	2.4	Use Case Diagram for Pie Chart	37
	2.5	Use Case Diagram for Generating Separate Files	38
3	3.1	System's Activity Diagram	42
	3.2	System's Activity Diagram (continued)	43
	3.3	Database Screenshot	44
	3.4	Initial work	49
	3.5	Second Step	50
	3.6	Step no.3	51
	3.7	Frequency Feature	52
4	4.1	DFD for Data Input module	53
	4.2	DFD for Database container module	54
	4.3	DFD for Result Module	55
	4.4	DFD for Frequency module	56
	4.5	DFD for Graphical module	57
	4.6	DFD for Data separator module	58
5	5.1	State Transition Diagram for Browse file	61
	5.2	State Transition Diagram for Check file	61
	5.3	State Transition Diagram for Pie Chart	62
	5.4	State Transition Diagram for Frequency	62
	5.5	State Transition Diagram for Separate files	62
6	6.1	Waterfall model	65
8	8.1	'That' Condition	80
	8.2	'Comma' Condition	80
10	10.1	Graphical representation of Result Table	90

## LIST OF TABLES

CHAPTER NO.	TABLE NO.	NAME OF THE TABLE	PAGE NO.
7	7.1	Estimated time of Project Completion	73
	7.2	Estimated Vs. Actual time of Project completion	74
10	10.1	Result Table	89



## **1. Introduction**

Chapter 1 is the “Introduction”; in this chapter we have given an overview of our project describing the inputs and outputs of the system, its domain, its market value. It also includes the history of the project and any prior work that is done on this system. It also consists of the tools and technology used to build the project.

### **1.1 What is Sentiments Analysis?**

Sentiments Analysis is also known as sentiment mining or opinion/sentiment extraction. Another alternative term is opinion mining, as it derives the opinion, or the attitude of a speaker. A common use case for this technology is to discover how people feel about a particular topic. It is the area of research that attempts to make automatic systems to determine human opinion from text written in natural language. It identifies the viewpoints defined within a certain text. The outcome anticipates the polarity of text under observation, i.e. positive, negative or neutral.

In simple words, it is the detection of attitudes behind words. Sentiments Analysis of natural language is a large and growing field these days. Sentiment classification distinguishes whether people like or dislike a specific product from their reviews.

The basic idea of Sentiment analysis is to understand how people respond to an ad campaign, product release, blog post etc. Was the product review positive or negative? Was the customer satisfied or unsatisfied? Based on a sample of tweets or comments, how are people responding to a certain ad campaign/product release/news item? How have bloggers' attitudes about the president changed since the election?

Sentiment Analysis determines sentiment on a variety of levels. It will score the entire document as positive or negative, and it will also score the sentiment of individual words or phrases in the document. Because Sentiment Analysis can track a particular topic, many companies use it to track or monitor their products, services or reputation in general.

The accuracy of Sentiment Analysis can be measured in many ways, but the most common way is to score accuracy in comparison to a human. Humans can only agree on whether or not a sentence has the correct sentiment, 80% of the time. There are a few major challenges for an engine analyzing text for sentiment.

One of the biggest issues is that it has trouble understanding irony. Even humans have trouble with someone who is being sarcastic. It is one of the most common mistakes an analysis system makes when trying to analyze text for sentiment.

Other problems can occur when words have multiple definitions. This happens quite a lot of times, because there are many words that are dual natured.

Sentiment analysis is a difficult task. The difficulty increases with the complexity of opinions expressed. Product reviews are difficult, books/movies/art/music reviews are more difficult, Policy discussions and indirect expressions of opinion more difficult and Non-binary sentiment (political leanings etc) is extremely difficult.

## **1.2 Goals and objectives**

Sentiment Analysis allows a person or an organization to get an opinion on reviews on a global scale. The main advantage is the speed. On average, humans process six articles per hour against the machine's throughput of 10 per second.

Marketers have been waiting for new ways to measure social media engagement. Measuring the attitude of a consumer towards a brand, through sentiments analysis is a method that is gaining popularity these days. For many marketers, the number of Facebook Likes and Twitter Follows they attract demonstrates their social standing. That's where sentiment analysis plays its vital role, as it adds meat to the bones of Likes and Follows.

Businesses are using the power of social media to gain a better understanding of their markets by identifying new trends for their product development team, to protecting their brand image and improving marketing campaigns and overall customer experience.

Sentiment analysis can help you with your business:

Get data from social media to understand attitudes, opinions and trends, and manage your online reputation.

Improve customer satisfaction by understanding customer needs and advice next best actions.

Create customized campaigns and promotions that coordinate with social media participants.

Identify and target the primary influencers within specific social network channels and approach them with unique offers.

### **1.2.1 Applications in Business**

Sentiment Analysis allows business to focus on:

Marketing intelligence

Product and service improvement

To understand the voice of the customer as expressed in everyday communications.

New Product Perception

Brand perception

Reputation Management

### **1.2.2 Applications in Business Intelligence**

Question: “Why aren't consumers buying our phone?”

We know the data: price, specifications, competition, etc.

We want to know subjective data

Misperceptions are also important

### **1.2.3 Political Sentiment Analysis**

Analyzing trends, identifying ideological bias and targeting advertisers, etc.

Public opinion– Attitudes to policies, parties, government agencies and politicians.

Policy-making– Arguments informing discussions between representatives.

Informal or formal environments.

### **1.3 System statement of scope**

The charm of this project is that it is sentiments analysis in Urdu language. In Urdu, there has been no significant work done on this topic till now. Whatever work that is done in Urdu has not been published or disclosed for general public.

Basic use of sentiments analysis in Urdu can be:

To analyze comments on Urdu websites

For Urdu newspapers to analyze reviews and political statements

#### **1.3.1 System Description**

Sentiment analysis simply concerns with Natural Language Processing. Natural language processing (NLP) is a field of Computer Science, Artificial Intelligence and linguistics that deals with the interactions between computers and Natural Language. [1]

The techniques used are:

- **Data Mining:**

Data mining is also known as knowledge discovery. It is a process in which data is examined from different point of views and angles, and then summarized into useful information. The information is of that type, which is used to increase revenue and cut costs.

- **Machine Learning:**

Machine learning is a branch of artificial intelligence, which involves the development and study of systems that can learn from data.

#### **1.3.2 Our Approach**

The basic problem is that the language tools are not compatible with Urdu language, therefore a meeting was held at the end of May 2014 with Sir Dr Sohail Sattar and Dr.Tafseer Ahmed (DHA Suffa University, Karachi) in which it was decided that a Rule-Based Approach will be followed and coding will be done in C# (Visual Studio) because Language Tools (Weka/Rapid Miner) don't give Urdu Support till date.

#### **1.3.3 Major inputs/processing functionality and outputs**

According to our approach, we are dealing with simple positive and negative adjectival phrases, phrases combined with polarity shifters, sentence separators, as well as interrogative sentences.

Our focus is to work only on sentiment classification for product reviews. Our domain is Samsung Product reviews/comments. Reviews can be good or bad, but it is seldom

neutral. Our system classifies the reviews as positive, negative or neutral based on the contextual sentiment orientation of the words.

### 1.3.4 Data Collection

Initially we took comments/reviews of people about Samsung Products in Roman Urdu and transliterated into Urdu script using "Google transliterate Urdu".

Examples of Different types of sentences:

- |                     |  |
|---------------------|--|
| 1) Simple positive  | -بہت ہی اچھا موبائل ہے                       |
| 2) Simple negative  | -یہ فون بہت برا اور ناکارہ ہے                |
| 3) Polarity reverse | -مجھے یہ فون پسند نہیں آیا                   |
| 4) Neutral          | سامسنگ فون اچھا ہے، کبھی خرابی بھی ہوجاتی ہے |
| 5) Interrogative    | کیا یہ فون اچھا ہے ؟                         |

### 1.3.5 Creating Data Dictionary

Urdu language shows exceptional grammatical features. For this, we use a lexicon-based approach. As there is no such lexicon available for this language, we created it manually i.e. the development of Urdu word Dictionary for a detailed analysis of adjectival phrases in Urdu text. The key point to be noted here is that data dictionary for Urdu has not been published by anyone, so this is a unique and noticeable task performed by us.

In this words dictionary/data base, we have two fields, i.e. of positive and negative words. By the time we worked upon this project and tested different conditions, we kept modifying our dictionary for better usage.

## 1.4 System context

This system that is Sentiments analysis is very useful for business and marketing nowadays e.g., tweets, blogs, social media, product reviews, political reviews, movie reviews, book reviews etc. Even a common person such as an actor, an artist, or a cricketer can get benefit from Sentiments analysis by simply knowing that how they can inspire the common public? How people think about them? Or how they can improve themselves? This system is used to understand the voice of the customers in

everyday communication that is what are the opinions of the people about any product or brand or any current issues?

As banker's appetite for trading lend is slowly starting to develop or expand, competition is increasing and every perspective of advantage needs to be considered to win business. So, what separates the star performers from the stragglers? Clearly, wisdom and strong credit disciplines are at the center of helping borrowers make smarter pecuniary decisions, but consider what makes your greatest correlation managers or senior credit offers victorious? Certainly, highly skilled individuals with vast experience who purchase or leverage awareness of their clients prosper at developing and maintaining relationships, structuring deals, evaluate and investigate credit requests and implementing or accomplishing complicated transactions. So, at the center of any successful organization is the ability of its individuals to influence information to get competitive understanding which can be turned into profitable business. So this can be done through sentiment analysis system.

Consider how this can help bankers improve credit decisiveness, underwriting, grow and pricing, enhancing their overall competitiveness encapsulating hard to quantify features of a firms' essentials such as insight into a borrower about the sincerity of a company's management, legal issues, deliberate direction, or changes in managerial encouragement, competitors, customers, industry growth potential, labor markets, regulation and product ratings based on unfiltered view of customer satisfaction as well as the latest news all at unprecedented speeds.

#### **1.4.1 Use of the system in Business**

Sentiments analysis is very helpful for determining the customer's reviews in business and marketing intelligence, so that a company can improve their service or product.

Business questions and Sentiment analysis answers

Social data allows you to comprehend

These are some of the questions that come in the mind of business workers to know the review of their product:

How do people feel about my product or brand?

How are people responding to our movements or product inaugurate?

Is there a way to predict the outcome of a campaign or event so that we know how we should invest in marketing?

Why aren't customers buying our product?

How to keep track of brand status or reputation?

The solution of these questions would be the use of Sentiment analysis system. Sentiment Analysis and social data enables business decision makers to understand consumer attitudes and behaviors more than ever before. Since social data is spontaneous and user generated, mining this data and categorizing it allows companies to understand intelligence around consumers' feelings towards movement, campaigns, content, or products.

How can it be used for business intelligence?

Social data allows you to comprehend information circulating around the whole industry. Sometimes, it can be hard to find out valuable information in surveys from people who didn't buy your product or service. You can use sentiment analysis to search the digital space for opinions, reviews, posts, and tweets around the competition. This is user generated understanding that you can collect for product positioning.

### **1.4.2 Prospects of the system in market:**

Market sentiment is the general convincing attitude of investors as to predict price development in a market. This attitude is the collection of a variety of fundamental and technical factors, including price history, economic reports, seasonal factors, and national and world events.

For example, if investors expect upward price movement in the stock market, the sentiment is said to be bullish. On the contradictory, if the market sentiment is bearish, most investors expect downward price movement. Market sentiment is usually considered as a contrarian indicator: what most people expect is a good thing to bet against. Market sentiment is used because it is believed to be a good predictor of market moves, especially when it is more extreme. Very bearish sentiment is usually followed by the market going up more than normal, and vice versa.

Market or business sentiments are monitored with a variety of methods such as the number of increasing stocks is compared with number of decreasing stocks. For calculating market sentiment one sentiment analytical tool is used. For example one tool extract movement on stock exchange and validly called market sentiment other tool extracts the news and media information based on their polarity.

Investors are also measures market sentiment through the use of news analytics, which include sentiment analysis on data about companies.

#### **1.4.2.1 Social marketing:**

In Social marketing responses and feedback through Facebook, YouTube, Twitter, and other platforms are analyzed.

Here are a few measures to calculate the response of the users:

Amplification rate – tracks the rate at which your audience shares your content

Applause rate – reveals the rate at which your audience provides positive feedback through share and likes

Audience growth rate – measures the rate of social network growth over a period of time

Visitor frequency rate – compares new and returning followers

The power of social media has given consumer voice to express their views. If any product gets likes, positive comments and shares it is obvious that it is popular among people.

How do brand managers and online marketers understand the diverse customer sentiment with regards to their brand online? What tools do they use to uncover these feelings from their online consumers, and most importantly how can this help to design a strategy to solve any problem and getting positive results?

Quality metrics include opinions, feelings, attitudes, the quality of shares, comments, re-tweets, replies, ratings or conversations, as well as the overall quality of engagement over time. The benefit of this analysis is that it can help in learning about



positive, negative or indifferent aspects of your brand being shared online and how to react to it.

The overall quantitative analytics packages are hugely important, but used in isolation can be highly misleading in terms of online brand engagement.

#### **1.4.3 Use by common people:**

Individuals can also benefit from sentiment analysis, whether they are a brand holder or they just want to know what people think of them and how popular they are. Actors, celebrities, sportsmen, authors and all other popular individuals can definitely benefit from the idea of Sentiments analysis. They can know how they inspire the common public and how people react to their certain move or activity and which attracts peoples; attention towards them. An ordinary man for example a blogger can also benefit from sentiment analysis.

#### **1.4.4 Strategic issues:**

The essential issues in sentiment analysis are to identify how sentiments are expressed in text and whether the expressions indicate positive (favorable) or negative (unfavorable) opinions toward the subject. In order to improve the accuracy of the sentiment analysis, it is important to properly identify the semantic relationships between the sentiment expressions and the subject. By applying semantic analysis with a syntactic parser and sentiment lexicon, good precision can be achieved.

Sentiments analysis is an area where we're seeing a lot of innovation in the industry. At a fundamental level, it's hard to understand how people feel about your brand in short reviews or comments. Getting machines to understand how people feel, their opinion and what they may say next about your company is critical to brands but is something that can give you some idea. So you can see all sorts of different technology out there that's trying to solve this puzzle which is sentiment analysis.

### **1.5 Theoretical Background**

Sentiment analysis is basically the attempt to extract the emotion or feeling of a body of text. The field of sentiment analysis and opinion mining usually also involves some form of data mining to get the text.

For sentiment analysis it is important to keep up technologies and solutions that help us to discover business value in opinions, feelings, and attitudes in social media and news. And further it also helps in the growth of business. Running a successful business is not possible when your information is not of a good level. In this world of globalization and socialization, an awareness of market trends is a vital component to making the right choices. So it is necessary to know customer's opinion about your product by using sentiment analysis system.

Like other languages, Urdu websites are becoming more popular, because the people prefer to share opinions and express sentiments in their own language. Sentiment analyzers developed for other well-suited languages, like English, are not workable for Urdu, due to their scripted, morphological, and grammatical differences. As a result, this language should be studied as an independent problem domain.

Urdu, the national language of Pakistan is fast losing its importance in this era of technology, it's our National language and a source of pride and honor for us it is the language which united the Muslims in the sub-continent.

Instead of crying over what might have gone wrong let's look at things which might be useful to promote the language of Indus civilization.

Our local language is Urdu then why we are giving much importance to English. Japan, China, Germany, Korea and Russia they all are well developed countries but they do not know English and if you want to work in these countries then you have to learn their local languages, development does not means that everyone in your country is communicating in English, development can be achieve by generating an environment of local language.

In English great advancements have been made in the field of Sentiment Analysis but Urdu is unfortunately still lacking behind. Urdu has become a source of expression, feelings, thoughts and aspiration. That is the reason we chose this project to promote our national language Urdu. Sentiment analysis has been implemented in other languages but not in Urdu, so we are trying to give Urdu a place in society and it is hope that Urdu would find its place in society within short period of time.

### **1.5.1 Background material provided**

When we started to work upon this project, we went through some research papers, books and web links:

Prabu Palanisamy, Vineet Yadav and HarshaElchuri [2] describes a lexicon based approach for discovering sentiments. They used preprocessing steps such as stemming, emoticon detection and normalization, exaggerated word shortening and hash tag detection. After the preprocessing, the lexicon-based system classifies the tweets as positive or negative based on the contextual sentiment orientation of the words.

G.Vinodhini and RM.Chandrasekaran [3] describes the Sentiment analysis involves in building a system to collect and examine opinions about the product made in blog posts, comments, reviews or tweets. Sentiment analysis can be useful in several ways. For example, in marketing it helps in judging the success of an ad campaign or new product launch.

III. Afraz, Aslam and Martinez [4] describes how Adjectival Phrases are combined with Polarity Shifters, and Conjunctions to make sentiment expressions in the opinionated sentences. These sentiment expressions are labeled as Senti Units. They are using a grammatically motivated approach which employs a Sentiment Annotated Lexicon (explanatory notes on Urdu words) are using shallow parsing based chunking. The classification algorithm works in combination with a sentiment-annotated lexicon of Urdu words.

IV. Afraz, Aslam and Martinez [5] describes the structure and construction of sentiment-annotated Urdu words based lexicon as a component of a sentiment analysis model developed for Urdu text. This approach recognizes the subjective entries in the lexicon through their two attributes; i.e. orientation (either positive or negative) and intensity.

The foremost task is data collection (Samsung Mobile Phone Comments). We took comments/reviews of people about Samsung Mobile Phones in Roman Urdu and transliterated into Urdu script using "Google transliterate Urdu".

## **1.6 Technology & Tools**

### **1.6.1 Technology:**

The approaches or technologies to perform sentiment analysis task are discussed here:

#### **“Bag of Words” Model**

This model focuses completely on the words, or sometimes a string of words, but usually pays no attention to the context. The bag of words model usually is a large list, a better version of a dictionary which contains words that carry positive and negative sentiment. These words each have their own “value” when found in text. The values are all added up and the result is calculated. The equation to add and derive a number can vary, but this model mainly focuses on the words and makes no serious attempt to actually understand basic language fundamentals.

#### **Natural Language Processing**

This model actually understands the sentences structures, context, and is more focused on the succession of a string of words. Usually, this structure requires the machine to have understanding of grammar principles. To do this, Natural Language Processing (NLP) techniques are used to tag parts of speech, named entities, and more, in order to actually understand the “language” of the text, and not just look for target words.

Since In Urdu we don’t have support on NLP tools hats why we are using bag of words model in our project which focuses completely on the bag of words i.e. dictionary and we used this approach by making the data dictionary.

#### **What is artificial intelligence?**

Artificial intelligence (AI) is the intelligence of a machines or software. It is a field of study that focuses on the goal of creating intelligence, whether in emulating human-like intelligence or not. Many researchers and textbooks define this field as "the study and design of intelligent agents", where an intelligent agent is a system that perceives its environment and takes actions that maximize its chances of success.

The central goals of AI research include reasoning, knowledge, planning, learning natural language processing (communication), perception and the ability to move and manipulate objects.

#### **What is Machine learning?**

Machine learning is a scientific/engineering discipline that deals with the construction and study of algorithms that can learn from data. Such algorithms operate by building a model based on inputs and using that to make predictions or decisions, rather than following only explicitly programmed instructions.

Machine learning can be considered a subfield of computer science and statistics. It has strong ties to artificial intelligence and optimization, which deliver methods, theory and application domains to the field. Machine learning is employed in a range of computing tasks where designing and programming explicit, rule-based algorithms is infeasible.

### **1.6.1.5 What is Natural Language Processing (NLP)?**

Natural Language Processing gives machines the ability to read and understand the languages that humans speak. An efficient and powerful natural language processing system would enable natural language user interfaces and the acquisition of knowledge directly from human-written sources. Some straightforward applications of natural language processing include text mining and machine translation.

A common method of processing and extracting meaning from natural language is through semantic indexing.

#### **Rule-based Approach**

Rule-based systems are used as a way to store and manipulate knowledge to store information in a useful way. They are often used in artificial intelligence applications and research.

A classic example of a rule-based system is the domain-specific expert system that uses rules to make deductions or choices. For example, an expert system can guide a doctor choose the correct diagnosis of a patient's disease.

Rule-based systems can be used to perform lexical analysis to compile or interpret computer programs, or in natural language processing.

Rule-based programming used to derive instructions from a starting set of data and rules.

A typical rule-based system has four basic components:

- A list of rules

- An inference engine which sorts information and make decisions on the interaction of input and the rule base. The interpreter executes a production system program by performing the match-resolve-act cycle.
- Temporary working memory.
- A user interface or other connection to the outside world through which input and output signals are received and sent.

We have tried to use Rule Based Approach in our project

### **1.6.2 Tools**

There are so many tools used for data mining, two of which are Weka and rapid miner. We decided to make our project on Weka or rapid miner. But our project is in Urdu text, so we tried to make this tool to be able to support Urdu language but these tools did not support Urdu language. After few months of struggle, we decided to make our project in Visual Studio(C#) Windows Form which supports Urdu language, instead of using data mining tools. This is called a rule-based approach by creating database or data dictionary we used MS Access.

#### **1.6.2.1 Weka**

Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in programming language. Weka is free software.

Advantages of Weka include:

- free availability
- portability
- a comprehensive collection of data pre-processing and modeling techniques
- ease of use due to its graphical user interface

Weka supports several standard Data mining tasks. Weka provides access to SQL databases using database connectivity and can process the result returned by a database query.

#### **1.6.2.2 Rapid Miner**

Rapid Miner is a software platform developed by the company of the same name that provides an integrated environment for machine learning, data mining, text mining,

predictive analytics and business analytics. It is used for business and industrial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the data mining process including results information visualization, validation and optimization. Rapid Miner is developed on a business source model which means the core and earlier versions of the software are available under an OSI-certified open source license on Source forge. Rapid Miner provides data mining and machine learning procedures including: data loading and transformation (ETL), data preprocessing and visualization, predictive analytics and statistical modeling, evaluation, and deployment. Rapid Miner is written in the Java programming language. Rapid Miner provides a GUI to design and execute analytical workflows. Those workflows are called “Process” in Rapid Miner and they consist of multiple “Operators”. Each operator is performing a single task within the process and the output of each operator forms the input of the next one.

### **1.6.2.3 Visual Studio Windows Forms**

Windows Forms is a .NET Library that wraps the native Windows application programming interface in managed code in order to allow the development of desktop applications. Windows Forms lets you create Windows-based apps capable of displaying data, handling user input and deploying software easily and securely. Windows Forms can function as a simpler to use replacement for the Microsoft foundation Class Library, which was based on the C programming language.

Windows Forms offers managed libraries for the .Net Framework. Windows Forms can be used within development environments such as Visual Studio to produce applications that take information from user and display data.

The Visual Studio Windows Forms Designer lets users create applications by selecting controls with a cursor and placing them on a form.

Windows Forms lets users display tabular data through a control called the DataGridView. This tool allows data to be shown in rows and columns so that each individual piece of information is shown within its own cell. The tool lets users lock rows and columns in place, change the appearance of cells and display controls within them.

Data sources can be integrated into Windows Forms through the DataGridView controller. The BindingSource component connects to data sources and enables users

to bind data to controls, navigate between records, perform editing tasks and save changes to the original source.

#### **1.6.2.4 MS Access**

Microsoft Access is a computer application used to create and manage computer-based databases on desktop computers and/or on connected computers (a network). Microsoft Access can be used for personal information management (PIM), in a small business to organize and manage data, or in an enterprise to communicate with servers.

Microsoft Access stores data in its own format based on the Access Jet Database Engine. It can also import or link directly to data stored in other applications and databases.

Software developer can use Microsoft Access to develop application software, and "power users" can use it to build software applications. Access is supported by Visual Basic for Applications, and it can reference a variety of objects including DAO (Data Access Objects), ActiveX Data Objects, and many other ActiveX components.



## **2. Usage Scenario / User Interaction**

This chapter will provide the details that how the users can interact with our system and how they can use this system to perform Sentiments Analysis on Urdu text. Intended users of the system are identified, Uses Cases diagrams for the system along with their description are given. It also constitutes of special considerations for using this system including data requirements and software requirements.

### **2.1 User profiles**

Our System is a desktop application. It's a dashboard in which user doesn't have to navigate between pages. All the results can be found on a single page. It doesn't require a user to set up an account or to sign in.

Only one user at a time can use our system.

After entering this system user just have to:

Browse their text file.

When text file opens in the system click on the Check button to start checking the system and wait for some time as the system prepares the results.

Now system will provide results as follows:

Total number of comments in the text file

Positive and negative words in each comment along with their frequency in ascending and descending order.

- Separate result of each comment
- Number of positive comments, number of negative comments and number of neutral comments along with the questions.
- Percentage of negative, positive and neutral comments.
- Graphical representation of result in a form of a Pie chart.
- Overall result of the text as positive, negative of neutral.

## **2.2Intended Users**

If anyone wants to use this system it is necessary for them to have familiarity with Urdu.

- User can be a company which wanted to have sentiments analysis of their product reviews.
- Anyone who wanted to the sentiment analysis of their products from the comments of people on social media can find this system useful.
- Users can be a Urdu newspaper agency or a news TV channel.
- User can be any common person who is interested to extract sentiments from a given Urdu text.

## **2.3Special usage considerations**

Special Data and Software requirements for the proper functioning of the system are listed here.

### **2.3.1 Data requirements**

Collect the product reviews which you wanted to check by the system.

If these are in English translate them in Urdu or if they are in Roman script transliterate them in Urdu script.

Combine all the reviews together in a text file.

### **2.3.2 Software requirements**

Version of operating system should not be older than Windows 7.

Processor should be 2.0 GHz Core Duo.

At least 3GB Ram is required

Install Visual Studio 2010.

Install Microsoft Access 2007 on your system.

## 2.4 Use cases

### 2.4.1 How Overall System works?



Figure 2.1 Use Case Diagram for the working of the whole system

### 2.4.2 Basic Concept of the System

Basically the system emphasizes on determining the sentiment of the product reviews, comments, newspaper text etc. Firstly the user has to select a domain or a product to identify its sentiment. After choosing a suitable product, the user collects reviews, comments related to the product and analyzes sentiments/opinions about it through the system.

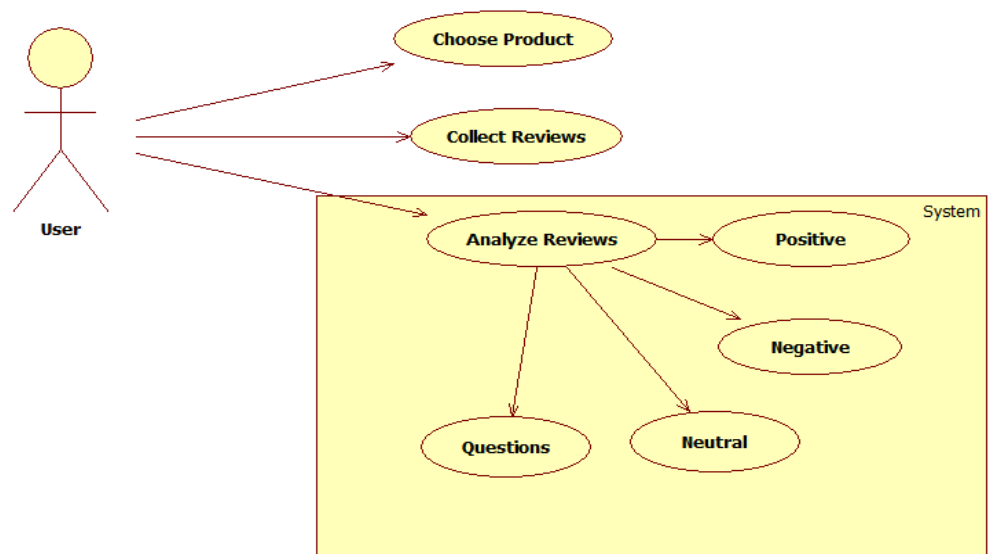


Figure 2.2: Use Case Diagram for the basic working of the system

### 2.4.3 Frequency of Positive and Negative Words

An overview for each component of subsystem 'n' is presented. omit the section if not applicable

The system first separates the positive and negative words from the comments file in a separate text boxes. It then checks the frequency of each positive and negative word that appears in each of the sentences in the text document. User can select the order of frequency to be ascending or descending.

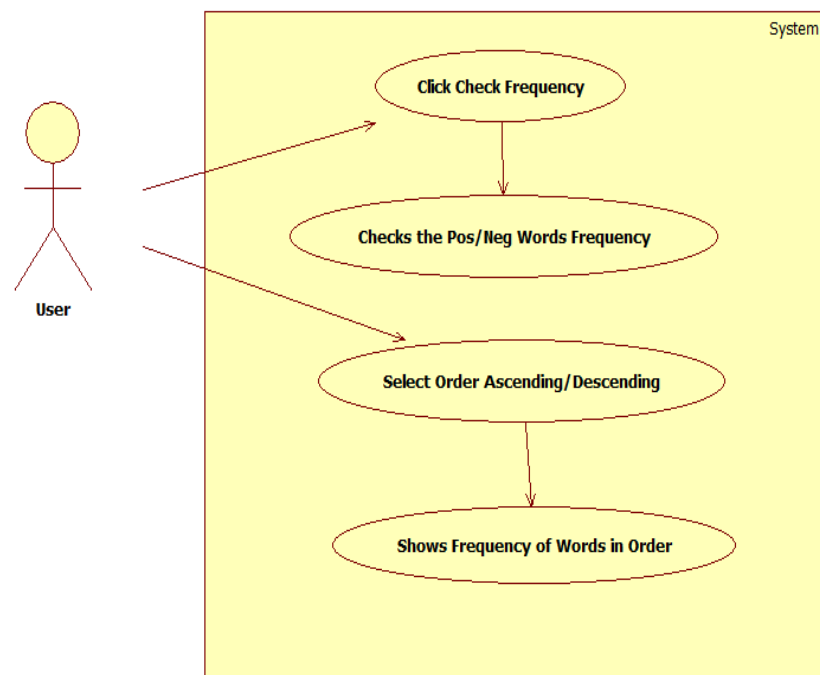


Figure 2.3: Use Case Diagram for Frequency Checking

#### 2.4.4 Show Pie Chart

After getting results of positive, negative, neutral and question sentences in percentage, we get a pie chart to illustrate numerical proportions clearly.

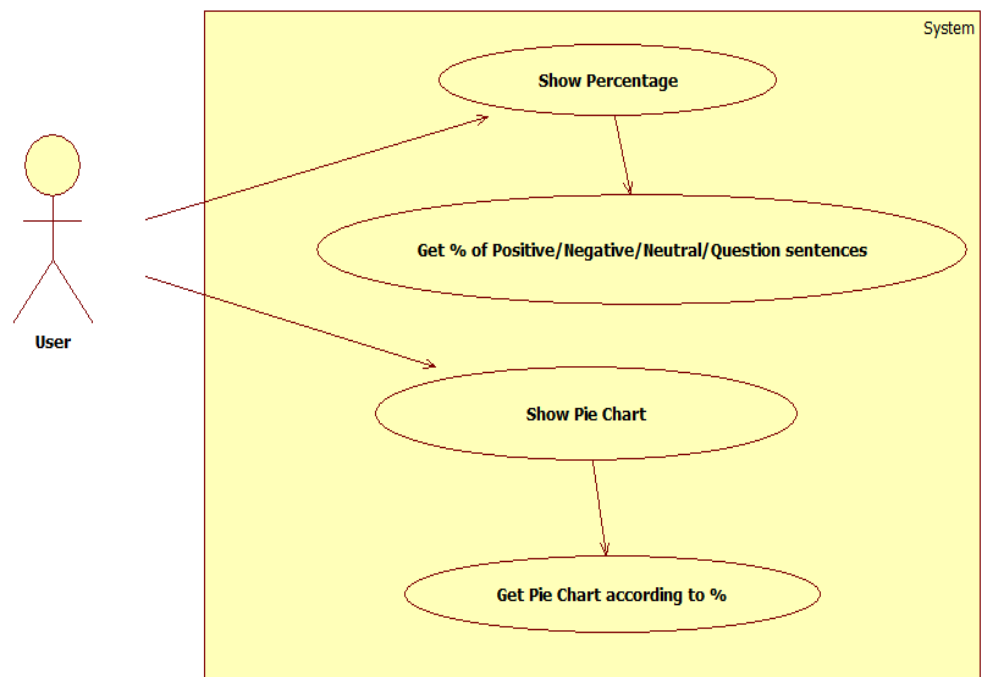


Figure 2.4: Use Case Diagram for Pie Chart

### 2.4.5 Generate Separate Files

Now we want to separate all the four types of comments/sentences. By clicking on the Generate Separate Files, we get Positive, Negative, Neutral and Questions files separately in the same directory from where we browsed our Comments.txt file.

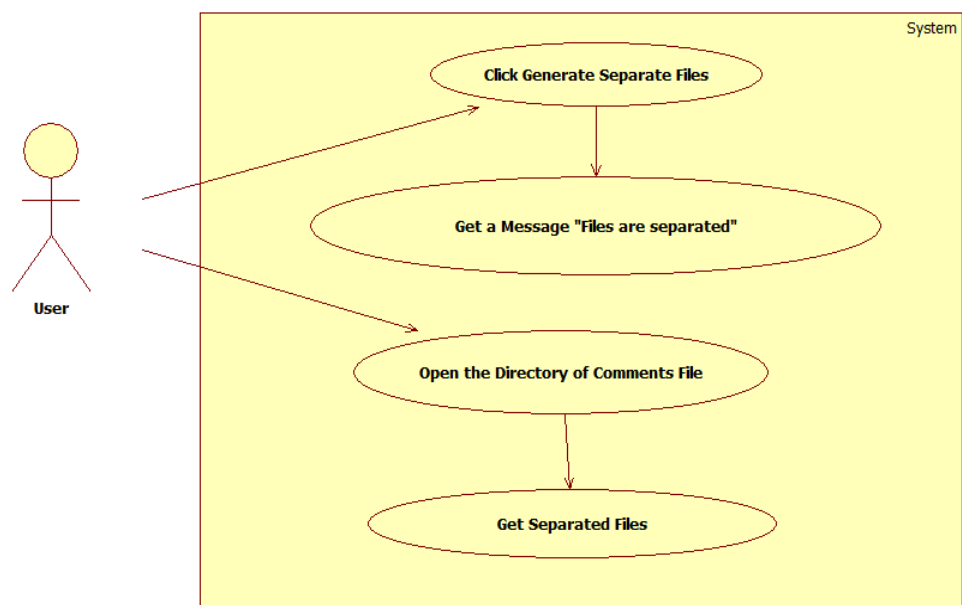


Figure 2.5: Use Case Diagram for Generating Separate Files





### **3. Functional and Data Description**

This chapter describes the architecture of the system. It contains the functionality of the system and type of information that flows through the system. It also consists of activity diagram of the system. Also contains information about the data objects and the sub modules of the system.

#### **3.1 System Architecture**

A context-level model of the system architecture is presented in this section. Overview is presented here and details in sub sections. Only brief overview of what this section will contain should be mentioned in this heading

##### **3.1.1 Architecture model**

Architecture of system is presented here. This includes detail description of system and system diagram

Detailed description of our system is:

This system is basically a Dash Board which is performing all of its tasks on a single platform. When the Check Button is clicked, the user is asked to browse his/her file. For our system, the file should be a text file which contains Urdu Sentences. On clicking 'OK', a message box pops up and shows the total number of comments in that selected file.

All of the comments are then displayed in the Rich Textbox. The comments are either separated by "-" OR "?" If the sentence separator is a question mark, it is identified as question and declared as Question in the list box. If sentence separator is a "-", then each word in that sentence is scanned in the Data Base. Our Data Base has two fields, i.e. Positive words and Negative words. If a word is found in the Positive field, then it is placed in the positive rich text box. Similarly, if a word is found in the Negative field, then it is placed in the negative rich text box.

For every word that is added to the positive/negative Rich text box, we increase the positive/negative count. Using these counts, we check:

If the document inputted by the user:

Contains only Positive comments

Result will be: Positive

Contains only Negative comments

Result will be: Negative

Contains only Neutral comments

Result will be: Neutral

Cases when Negative and Positive comments are present in document and no Neutral comments are there:

If  $\text{Positive} > \text{Negative}$

Result will be: Positive

If  $\text{Negative} > \text{Positive}$

Result will be: Negative

$\text{Positive} = \text{Negative}$

Result will be Neutral

Cases when Negative and Neutral comments are present and no positive comments are there:

$\text{Neutral} > \text{Negative}$

Result will be Neutral

$\text{Negative} > \text{Neutral}$

Result will be Negative

$\text{Neutral} = \text{Negative}$

Result will be Neutral

Cases when Neutral and Positive comments are present and no negative comments are there:

Positive > Neutral

Result will be positive

Neutral > Positive

Result will be Neutral

Neutral=Positive

Result will be Neutral

Cases when Neutral and Positive and negative all comments are present:

Then the one with the greatest percentage is taken as the final result.

Then we analyze and finalize the final result as positive, negative or neutral in the Result textbox. We calculate the percentages of the positive, negative and neutral separately and give its graphical representation through a pie chart.

Also, on clicking Go button, the frequency of words will be calculated separately for positive and negative words in ascending or descending order; as the user wishes.

On clicking the “Generate separate result files”, 4 separate files are generated for positive, negative, neutral comments, as well as a file with questions.

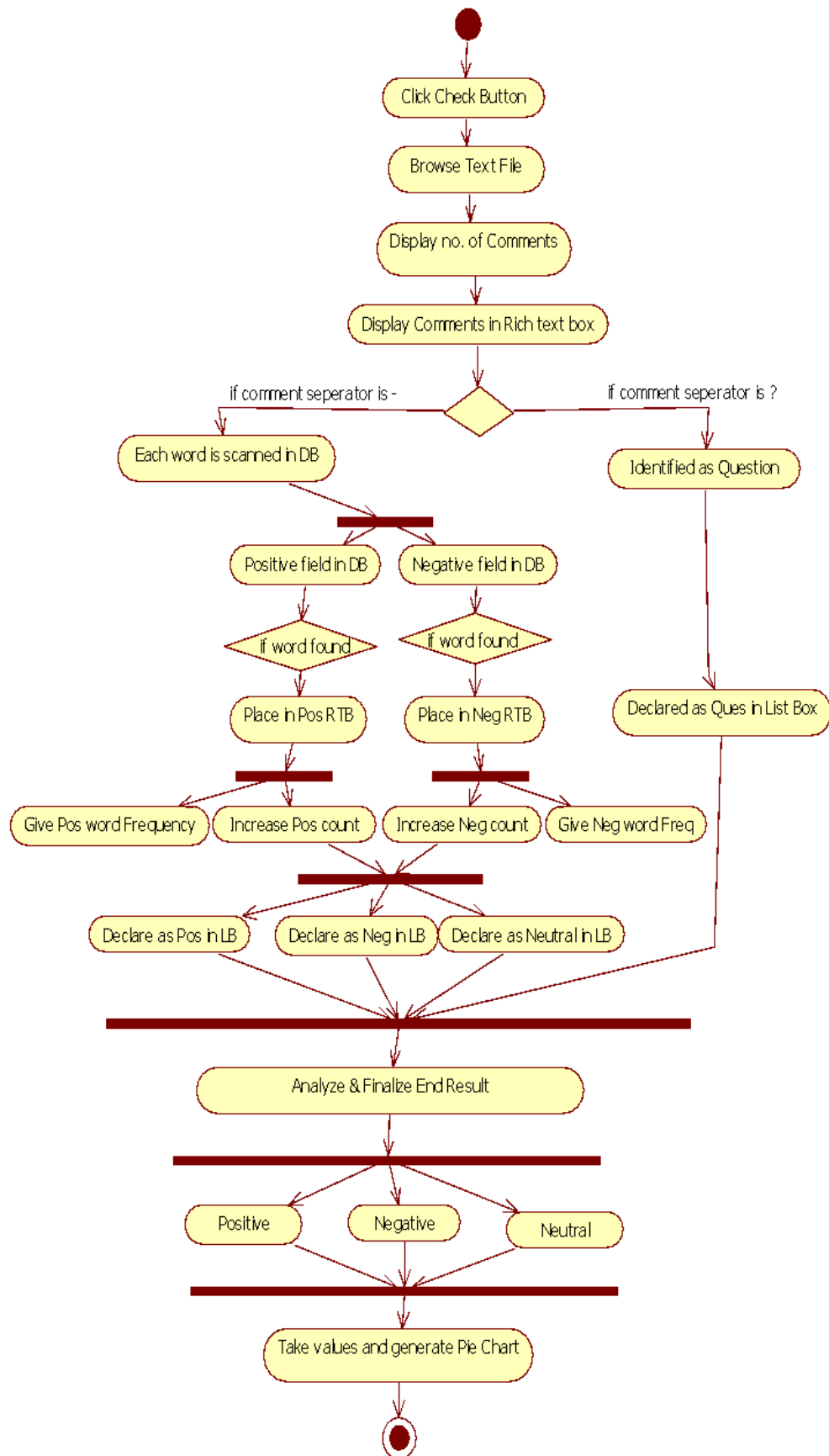


Figure 3.1

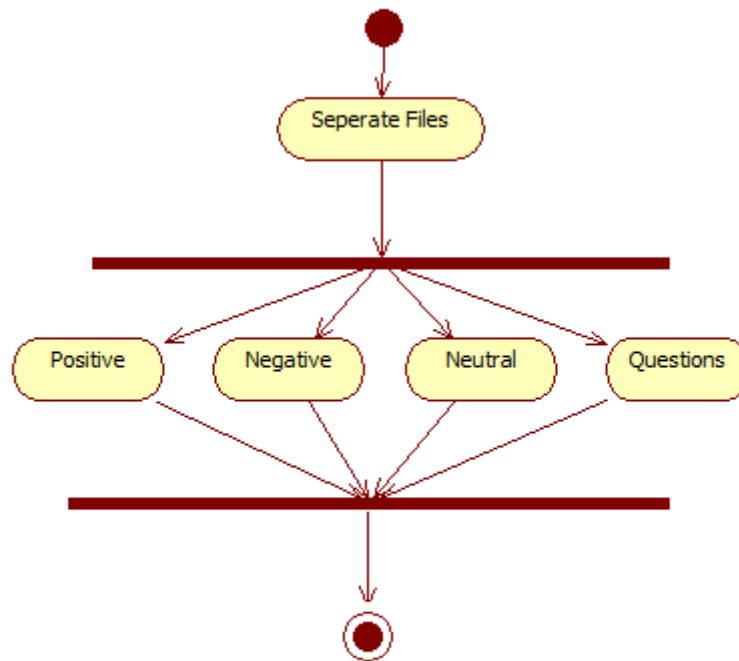


Figure 3.2

### 3.1.2 Subsystem/modules overview

The subsystems of our main system “Sentiments Analysis in Urdu Text” are:

#### **Data Input Module:**

It contains the data (Text document) inputted by the user.

#### **Database Container Module:**

Database module is the main module of our system. Database contains our Data Dictionary that is list of all the Negative words and Positive words are present in the database. All the inputted by the user is scanned and matched from the database to get the results.

#### **Result Module:**

This module calculates:

Positive and negative words in each sentence

Separate result of the each sentence.

Collective results for all sentences.

Total number of Positive, Negative and Neutral sentences and questions.

Total percentage of Positive, Negative and Neutral sentences and questions.

Overall result of the document whether the data inputted by the user is Positive, Negative or Neutral.

#### Frequency Module:

This module calculates the frequency of the all the positive words and all the negative words in ascending and descending orders.

#### Data Separator Module:

This module separates the results into 3 different files placing positive sentences in positive file, negative in negative file and neutral in neutral file.

#### Graphical Module:

This module shows the graphical representation through Pie Chart.

## 3.2 Data Description

The creation of lexicon, collection of different types of data and database creation are described here:



pos_words	neg_words
اچھی	شکایت
اچھے	قیاحتیں
آبدار	مایوس کن
آبداری	کھو
آب و تاب	پیچیدہ
آباد	وبال
اتفاق	بکواس
ایدی	مہنگا
آبرو	خبردار
اتحاد	بری
احتیاط	مسئلے
اٹل	ٹوٹ
اجاگر	مسئلہ
اجالا	خرابی
اجرت	برا
اجلی	یاگل
آرام	اٹک
آرام دہ	اٹکنا
آرزو	پچھتا
آرزومند	غلطی
اجملہ	دیکار

Figure 3.3 : Database Screenshot

### 3.2.1 Major Data Objects

We took comments/reviews of people about Samsung product in Roman Urdu and transliterated into Urdu script using "Google transliterate Urdu" which can be found at: [6]. We created our lexicon in Urdu, and collected sentences of different types which are described here:

### 3.2.1.1 Sentiment Lexicon Generator

The succeeding task is to prepare our data dictionary i.e. to identify nouns, verbs, adverbs, adjectives, positive and negative adjectives and intensifiers.

As a starting point we used [7] this dictionary. However, we found some problems with the lexicon. The problems are that they have used some equivalent words in both positive and negative lists and they have also used many incorrect words in their lists too.

We solved these problems by removing 1248 words from positive list and 2399 words from the negative list and added 61 words in positive list and 52 words in negative list. The total numbers of positive and negative words in their lists were 2607 and 4728 respectively.

### 3.2.1.2 Different types of sentences:

The different types of sentences that we collected for testing which is sample data are described here:

Simple positive

-بہت ہی اچھا موبائل ہے

In roman Urdu: Bohat he acha mobile hai.

In English: Very nice mobile

-بہت ہی اعلیٰ اور بہترین موبائل ہے

In roman Urdu: Bohat he aala aur behtreen mobile hai.

In English: Very great and best mobile

Simple negative

-یہ فون بہت برا اور ناکارہ ہے

In roman Urdu: Yeh phone bohat bura aur nakar hai.

In English: This phone is very bad and useless

Simple neutral

سامسنگ فون اچھا ہے، کبھی خرابی بھی ہو جاتی ہے -

In roman Urdu: Samsung phone acha hai, kabhi kharabi bhe ho jati hai.

In English: Samsung phone is good, sometimes it also malfunctions

Polarity reverse

مجھے یہ فون پسند نہیں آیا

In roman Urdu: Mujhay yeh phone pasand nahin aya.

In English: I did not like this phone.

یہ فون بالکل بھی برا نہیں ہے

In roman Urdu: Yeh phone bilkul bhi bura nahin hai.

In English: This phone is not bad at all.

Sentences joined by “مگر” and “لیکن” (“but”)

بہت اچھا ہے مگر اسکی بیٹری بہت کم چلتی ہے -

In roman Urdu: Bohat acha hai magar is ki battery bht kam chalti hai.

In English: It's great, but the battery is running low.

میں اسکی کارکردگی سے خوش ہوں لیکن کبھی کبھی یہ تنگ بھی کرتا ہے -

In roman Urdu: Main is ki kar kardigi say khush hoon lekin kabhi kabhi yeh tang bhi karta hai.

In English: I am happy with its performance but sometimes it bothers me.

If the first sentence is negative, the other will be positive. And if the first is positive, the second will be negative and overall impact of the full sentence will be Neutral.



Sentences joined by “کہ” (“that”)

یہ بات تو بالکل ٹھیک ہے کہ سامسنگ خراب فون ہے

In roman Urdu: Yeh baat to bilkul theek hai keh Samsung kharab phone hai.

In English: It's absolutely right that Samsung is a bad phone.

یہ بات غلط ہے کہ سامسنگ اچھا فون ہے -

In roman Urdu: Yeh baat ghalat hai keh Samsung acha phone hai.

In English: It is wrong that Samsung is a good phone.

Sentences joined by “کیونکہ” (“because”)

مجھے سامسنگ نہیں پسند کیونکہ یہ بار بار اٹکتا ہے -

In roman Urdu: Mujhay Samsung nahin pasand kyon keh yeh baar baar atakta hai.

In English: I do not like Samsung because it hangs repeatedly.

In these type of sentences, both the joined sentences will have the same polarity i.e. both of them will be positive or both of them will be negative. And the overall impact will be positive or negative.

Sentences joined by “،”

بالکل ٹھیک کہا آپ نے ، بہت بیکار ہے یہ فون -

In roman Urdu: bilkul theek kaha ap nay , bohot bekar hai yeh phone

In English: “You said exactly right, this phone is pretty useless.”

میں تو سامسنگ کمپنی اور اسکے موبائل کی فین ہو گئی ہوں ، ساری کمپنیوں میں یہ سب سے اچھی ہے -

In roman Urdu: main to Samsung company aur is kay mobile ki fan ho gaye hoon, sari companion main yeh sab se acha hai.

In English: “I have become a fan of Samsung, Samsung is best amongst all.”

In these types of sentences the second half of the sentence will decide the impact of the whole sentence. If the second part is positive, the whole sentence will have a

positive impact and if the second part is negative, the whole sentence will have a negative impact.

Interrogative sentences

کیا یہ فون اچھا ہے؟

In roman Urdu: Kia yeh phone acha hai?

In English: Is this phone good?

کیا سیمسنگ موبائل قابل خرید ہے؟

In roman Urdu: Kia Samsung mobile qabil e khareed hai?

In English: Is Samsung phone worth buying?

### **3.3 Major data objects**

Data objects and their major attributes are described.

### 3.3.1 System Interface Description

سیمنگ موبائل مارکیٹ میں آج بہترین فروخت ہوئے والے فونز ہیں

Check

	pos_words	neg_words
▶	good	bad
	nice	worst
	beautiful	مہنگا
	well	سکا پت

positive words: بہترین  
1

negative words:   
0

Result: positive

**Figure 3.3**

**The very first program for simple sentences. only one sentence can be given as input at a time.**



Figure 3.4

This is the next level of our system. In this, we have multiple sentences in the drop-down menu and calculate their result one by one.

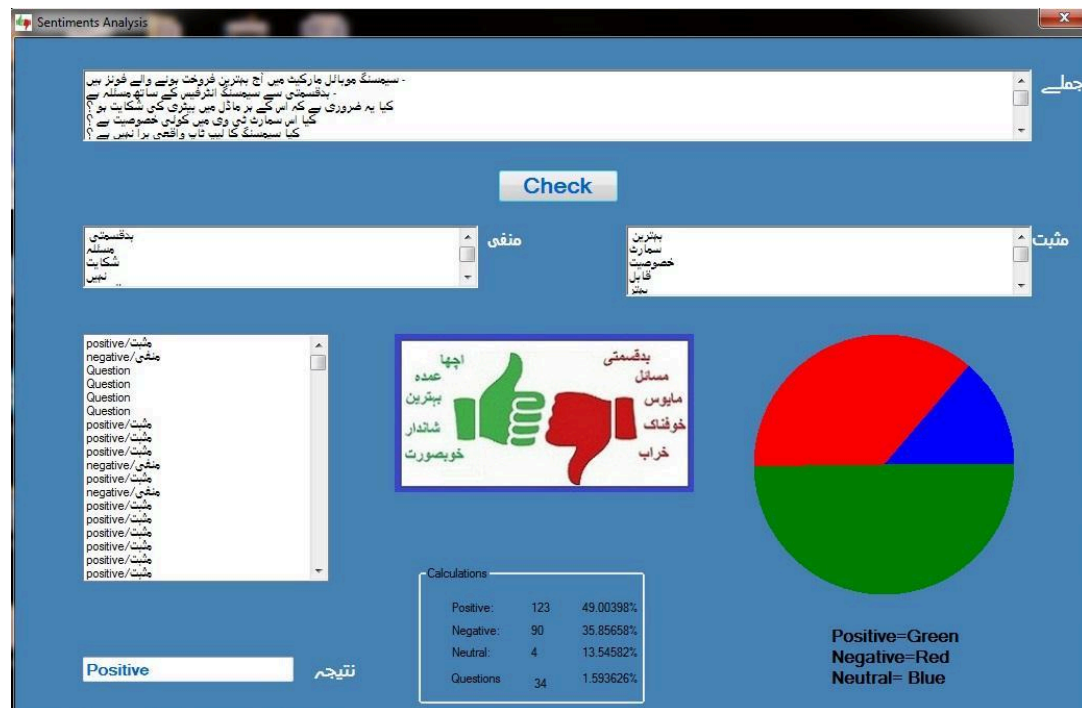


Figure 3.5

This is our third step. We give multiple sentences as input in a rich textbox and on checking our results, we get separate positive and negative words and we get separate result of each comment in the list box. It gives final collective result along with percentage and pie chart.

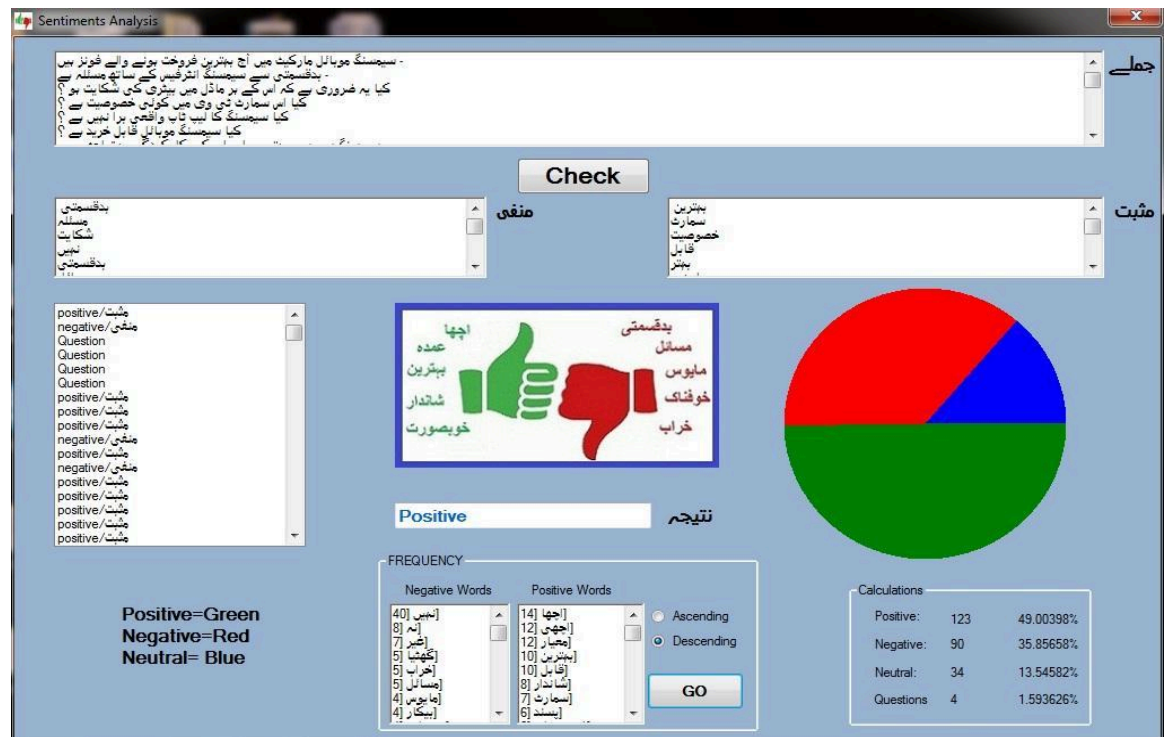


Figure 3.6

This is our fourth step. We give multiple sentences as input in a rich textbox and on checking our results, we get separate positive and negative words and we get separate result of each comment in the list box. It gives final collective result along with percentage and pie chart. We have now added the frequency feature, which can find word frequency for positive and negative words in ascending/descending order.

## 4. Subsystem/module Description

Subsystems are the group of interconnected and interactive parts that performs an important job or task as a component of a larger system. Our system compromises of the following six subsystems:

- Data Input Module
- Data Base Container Module
- Result Module
- Frequency Module
- Graphical Module
- Data Separator Module

### 4.1 Description for Data Input Module

This module is about inputting data in our system for its Sentiment Analysis.

#### 4.1.1 Subsystem scope

Here user will have to browse the targeted file which contains Urdu data. When the file is opens all its sentences will be seen in the text box. User will then press the check button to perform the sentiment analysis of given text

#### 4.1.2 Subsystem flow diagram

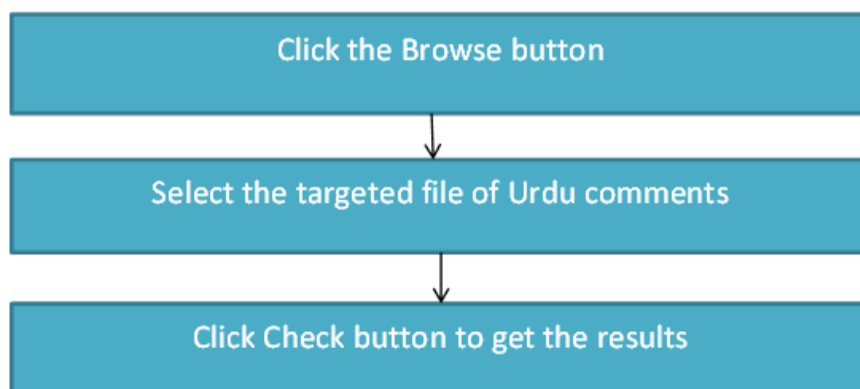


Figure 4.1: DFD for Data Input Module

## 4.2 Description for Database Container Module

This module contains the Database which is our Urdu Word Dictionary/Lexicon.

### 4.2.1 Subsystem scope

Document is broken down into sentences and sentences are broken down into words then each word in that sentence is matched with the Data Base. Our Data Base has two columns, Positive words and Negative words. If a word is found in the Positive field, then its positive count is increased. Similarly, if a word is found in the Negative field, then negative count is increased.

### 4.2.2 Subsystem flow diagram

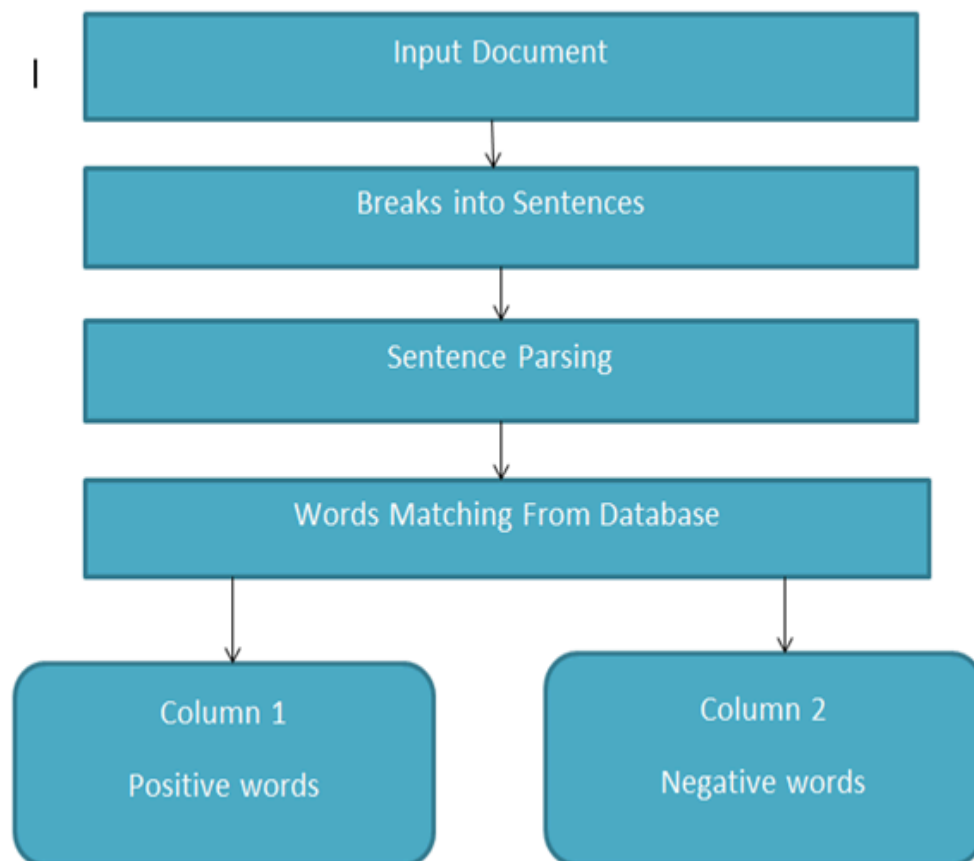


Figure 4.2: DFD for Database container module



### 4.3 Description for Result Module

This module gives the final result of the sentiment analysis of the given text.

#### 4.3.1 Subsystem scope

This module performs many actions. Firstly it gives the all positive and negative word in each sentence. Then by counting the positive word count and the negative word count as well as applying the polarity reversal condition it classifies each sentence as positive, negative or neutral. Then by calculating all positive, negative, neutral sentences and questions it gives the overall result of the document. It gives total number of Positive, Negative and Neutral sentences and questions as well as the total percentage of Positive, Negative and Neutral sentences and questions.

#### 4.3.2 Subsystem flow diagram

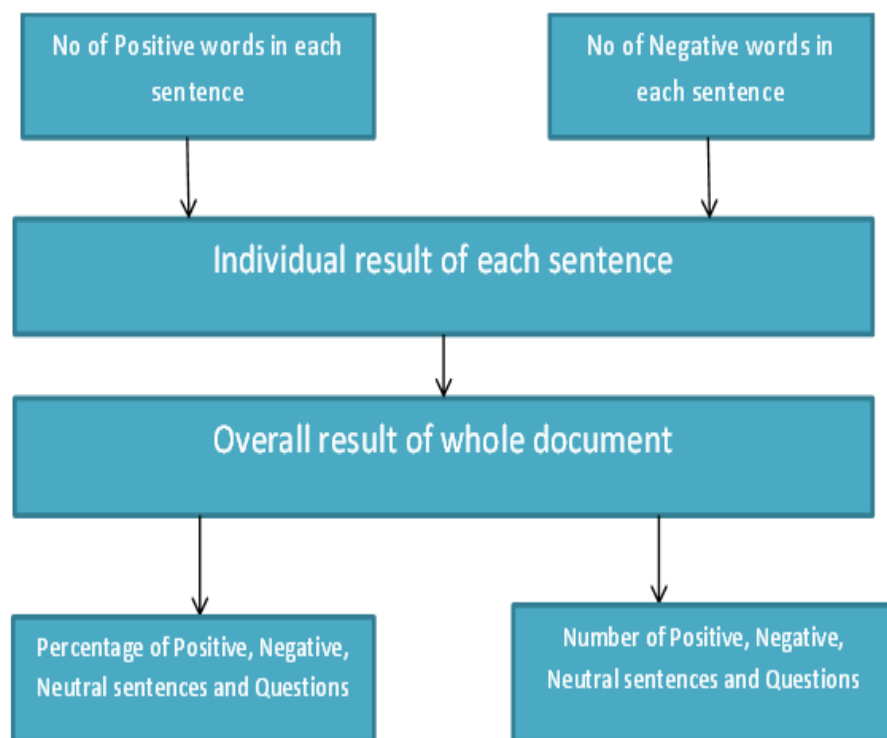


Figure 4.3: DFD for Result Module

## 4.4 Description for Frequency Module

This module is about inputting data in our system for its Sentiment Analysis.

### 4.4.1 Subsystem scope

Here user will have to browse the targeted file which contains Urdu data. When the file is opens all its sentences will be seen in the text box. User will then press the check button to perform the sentiment analysis of given text

### 4.4.2 Subsystem flow diagram

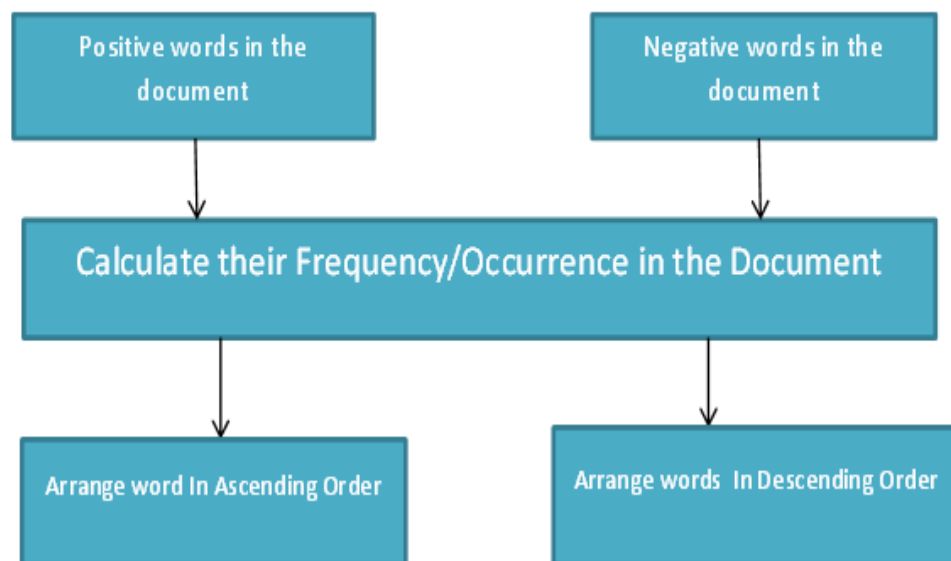


Figure 4.4: DFD for Frequency Module

## 4.5 Description for Graphical Module

This module gives the graphical result in the form of a Pie Chart.

### 4.5.1 Subsystem scope

It takes the percentages of positive sentence, negative sentences and neutral sentences from the result modules and according to them it generates a Pie Chart. The type of sentence with greatest percentage occupies the greatest area in the pie chart and the type of sentence with least percentage occupies the smallest area in the pie chart. Red indicates negative, green indicates positive and blue indicates neutral. By looking at the pie chart you can easily depict the overall result of the sentiments analysis.

### 4.5.2 Subsystem flow diagram

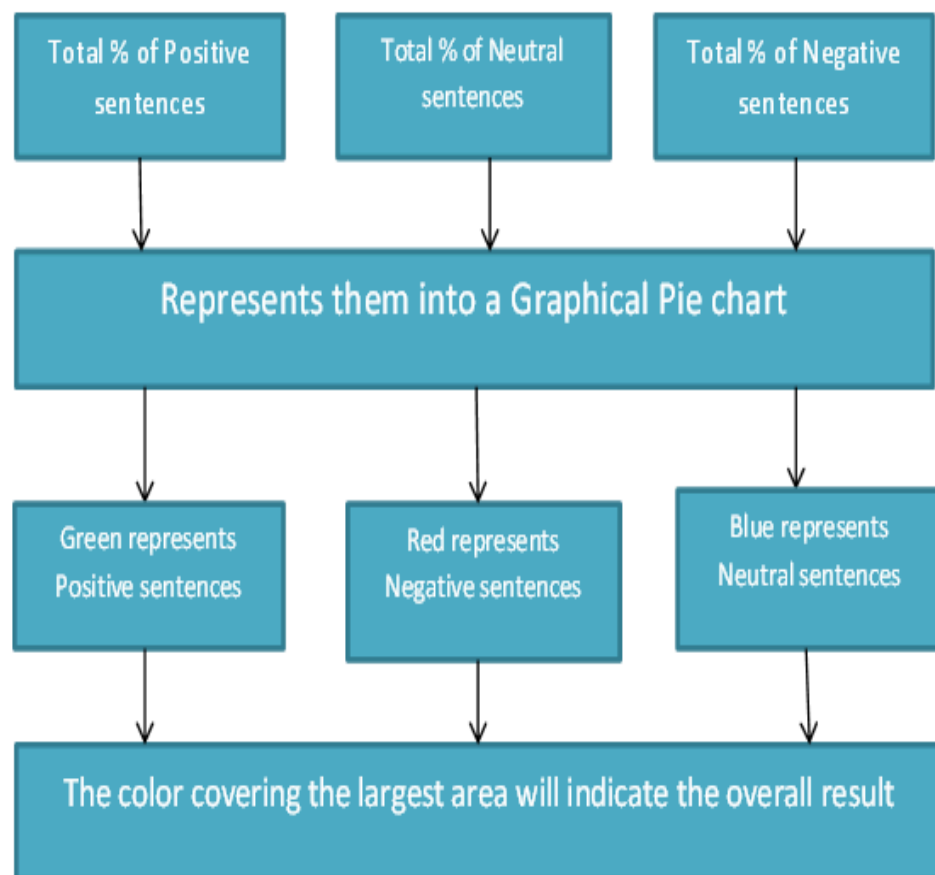


Figure 4.5: DFD for graphical module

## 4.6 Description for Data Separator Module

This module separates the results into 4 different files placing positive sentences in positive text file, negative in negative text file, neutral in neutral file and questions in questions text file.

### 4.6.1 Subsystem scope

After the calculation of overall total result the work of this module is to separate the positive negative and neutral sentences and questions. It generates four different text files each containing positive sentence, negative sentences, neutral sentences and questions. The files are generated in the same directory where the original text file which was inputted in our system is present. As well as calculating the sentiment analysis on the given Urdu text it also classifies it and separates them according to their category.

### 4.6.2 Subsystem flow diagram

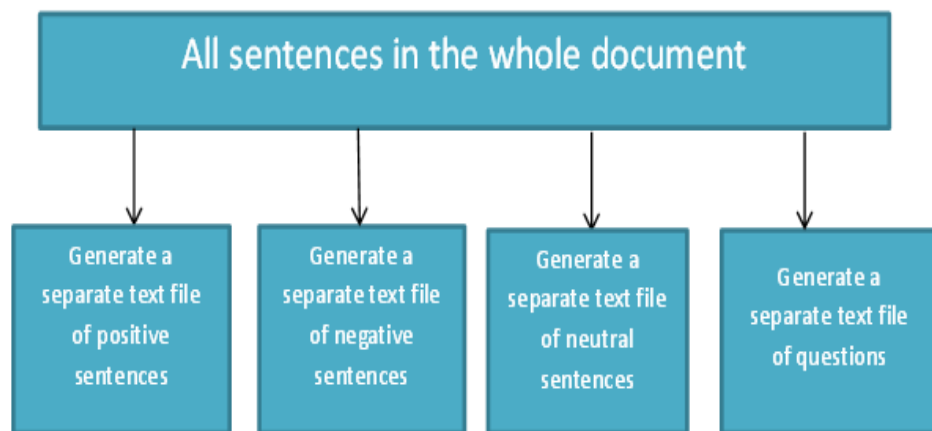


Figure 4.6: DFD for data separator module

## **5 Behavioral Model and Description**

Chapter 5 is “Behavior Model and Description”, this chapter discusses behavior of the system. Different states of the system are explained and event diagrams are drawn.

### **5.1 Description for System Behavior**

Sentiment Analysis aims to determine the attitude of a person, reader or writer, with respect to some topic or the overall contextual polarity of a document.

The major task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level, whether the expressed opinion in a document, a sentence or an entity aspect is positive, negative, or neutral.

The other subtasks include separate result of each sentence, overall result, calculation of results in percentage, graphical representation by pie chart, frequency of positive/negative words, generating separate files i.e. positive, negative, neutral and questions.

#### **5.1.1 Events/interrupts**

A listing of major events (control, items) that will cause behavioral change within the system is presented.

Browse comments file

Check file

Separate positive/negative words

Show separate results of each comment

Show collective result of the document

Show results in percentage

Show pie chart

Get frequency

Select order of frequency

Generate separate files

### **5.1.2 States**

A listing of states (modes of behavior) that will result as a consequence of events is presented.

File is browsed

The user browses the file containing training/testing data from his computer which then appears in the textbox.

File Checked

When the user clicks Check, all the comments in the file are checked separately, showing separated positive and negative words, result of each comment separately and then the overall the result.

Get Percentage

The system shows the results in percentage and then the graphical representation through pie chart.

Get Frequency

Get the frequency of all positive and negative words in ascending or descending order.

Separate Files are generated

This generates the separate positive, negative, neutral and questions file in the same directory from where you browsed your comments file.

## 5.2 State Transition Diagrams

Depict the overall behavior of the system.

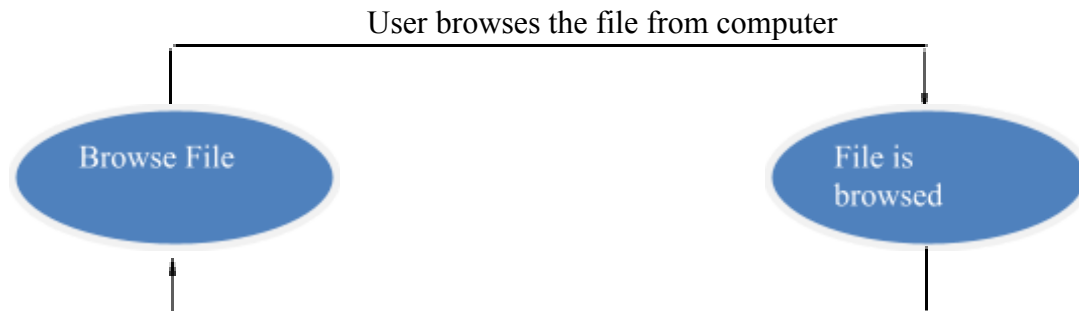


Fig 5.1: State Transition Diagram for Browse File

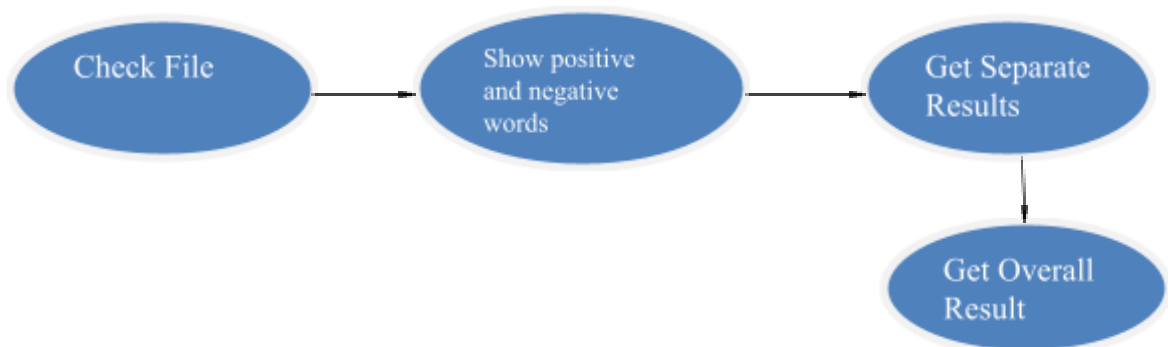


Fig 5.2: State Transition Diagram for Check File



Fig 5.3: State Transition Diagram for Pie Chart



Fig 5.4: State Transition Diagram for Frequency of Words



Fig 5.5: State Transition Diagram for Separate Files



## **6 System Prototype Modeling and Simulation Results**

If Chapter 6 is “System Prototype Modeling and Simulation Results, this chapter contains that the detail about the prototype you have designed before building the actual model. We have used the Waterfall model.

### **6.1 History of the Waterfall Model**

The “Waterfall Model” was first introduced by Winston Royce in 1970 in one of his articles, though he never used the word “waterfall.” He later presented this model to depict a flaw in a non-working model. So later on, this term was mostly used in writing about something that is often wrongly done in the process of software development – like a common malpractice.

### **6.2 What is Waterfall Model?**

The Waterfall Model was first Process Model to be known. It is also referred to as a linear-sequential life cycle model. It is easier to understand and use. In a waterfall model, each phase must be completed fully before starting the next phase. This type of model is basically used for the project which is small and there are no uncertain requirements. At the end of each phase, a review takes place to identify if the project is on the right path and whether or not to continue or discard the project. In this model the testing starts only after the development is complete. Phases do not overlap in this model. [8]

### **6.3 When should we use the Waterfall Model?**

This model is used only when the requirements are very well known, clear and fixed.

Product definition is stable.

Technology is understood.

There are no ambiguous requirements.

Ample resources with required expertise are available freely.

The project is brief.

Very less customer enter action is involved during the development of the product.

Once the product is ready then only it can be demoted to the end users. Once the

product is developed and if any failure occurs then the cost of fixing such issues are very high, because we need to update everywhere from document till the logic.

#### **6.4 Features of the Waterfall Model**

A Waterfall Model is easy to flow.

Every stage has to be done separately at the right time so you cannot jump stages.

Documentation is produced at every stage of a waterfall model that allows people to understand what has been done.

Testing is done at every stage.

The project requires the completion of one phase, before going ahead to the next one. Therefore, if there is an error in this software it will be detected during one of the initial phases and will be sealed off for corrections.

A lot of emphasis is laid on paperwork in this method as compared to the other newer methods. When new workers enter the project, it is easier for them to carry on the work from where it had been left. The newer methods don't document their developmental process which makes it difficult for a newer member of the team to understand what step is going to follow next. The Waterfall Model is a straight forward method and makes it easier for the person to know what stage is going on.

The Waterfall method is also well known amongst the software developers therefore it is easier to use. It is convenient to develop various software through this method in short period of time. [9]

## 6.5 How Waterfall Model is used in our Project?

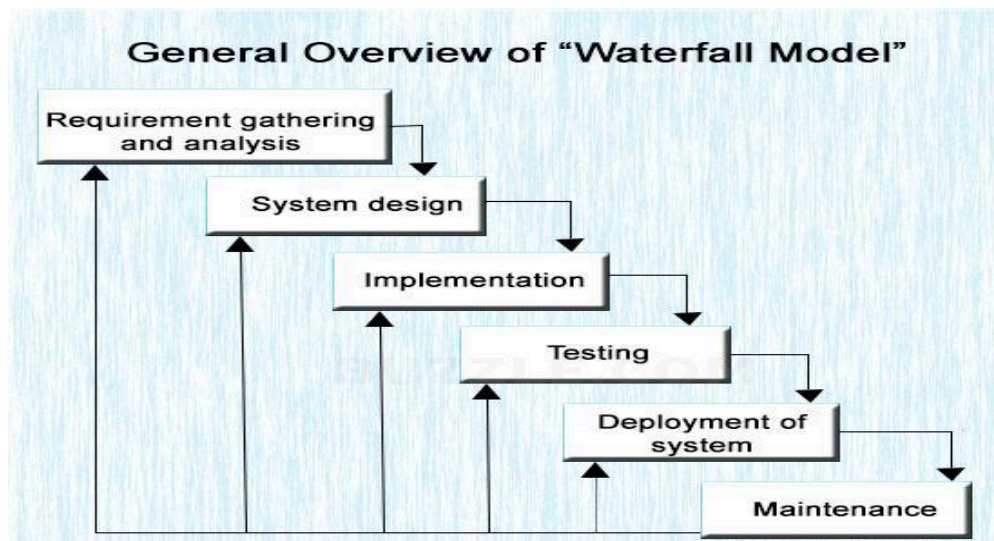


Figure 6.1: Waterfall Model

### 6.5.1 Requirement Gathering and Analysis:

All possible requirements of the system to be developed are gathered in this phase and these are documented in a requirement specification document. Requirements are a set of constraints that the end user expects from the system. The requirements are gathered from the end user, and are analyzed for their validity and the possibility of incorporating them. Finally, a requirement specification document is created which serves the purpose of a guideline for the next phase of the model.

This is the most crucial phase as any misinterpretation at this stage may give rise to validation issues later. The software definitions must be detailed and accurate with no ambiguities. It is very important to understand the customer requirements and expectations so that the end product meets its specifications.

The basic requirements of the system must be understood by software engineer, who is also called Analyst. All these requirements are then well documented and discussed further with the customer for reviewing.

In our project, in this phase we collected the requirements for the project i.e., all the requirements that the end user expects from the system.

### 6.5.2 System Design



The requirement specifications from first phase are studied in this phase and system design is prepared. System Design helps in specifying hardware and system requirements and also helps in defining overall system architecture.

The inter relation between the various logical modules is established at this stage. Algorithms and diagrams defining the scope and objective of each logical model are developed. This phase lays a fundamental for actual programming and implementation.

It is an intermediate step between requirements analysis and coding. Design focuses on program attribute such as Data structure, Software architecture, Algorithm details etc. The requirements are translated in some easy to represent form using which coding can be done effectively and efficiently. The designing needs to be documented for further use. The system design specifications serve as an input for the next phase of the model.

In this phase, we prepared our system design and algorithm for coding and designing. In the starting, our algorithmic construction was:

-Testing Data: Samsung Mobile Phone Comments

-Tracking the frequencies of adjectives with positive and negative connotations

foreach word in the sentence

if word is positive

positiveCount += 1

if word is negative

negativeCount +=1

// We reverse the polarity of a sentiment word whenever it is preceded by a negation.

If indexOf negation with a positive word is found

positiveCount -= 1

negativeCount +=1

If indexOf negation with a negative word is found

negativeCount -=1

positiveCount += 1

Result:

-Classification of the sentence or training data as positive or negative.

-If the frequencies are same, we have neutral data.

if positive words > negative words then

return positive

if negative words > positive words then

return negative

if positive words=negative words then

return neutral

### 6.5.3 Implementation



With inputs from system design, the system is first developed in small programs called units, which are integrated in the next phase. Each unit is developed and tested for its functionality which is referred to as Unit Testing. Unit testing mainly verifies if the modules/units meet their specifications.

Coding is a step in which design is translated into machine-readable form. If design is done in sufficient detail then coding can be done effectively. Programs are created in this phase.

In this phase, all software is divided into small modules because coding for small modules is easier rather than coding for the whole software. According to design, programmers do the code and make class and structure of whole software.

In the implementation phase, we break our algorithm and the whole code into smaller units i.e. we made separate conditions for all the Urdu text and developed separate programs for every condition, then combine all the separate conditions in a single program.

We constructed a separate program for “کہ” (“that”) condition which worked separately for many type of “کہ” sentences e.g.

یہ بات تو بالکل ٹھیک ہے کہ سامسنگ خراب فون ہے

یہ بات غلط ہے کہ سامسنگ اچھا فون ہے -

مجھے یہ کہتے ہوئے خوشی ہو رہی ہے کہ سامسنگ بہترین فون ہے -

اس میں ایک خرابی ہے کہ یہ بار بار بند ہو جاتا ہے -

But when we combined this code with all other conditions, then this program was not showing the correct result for some type of sentences e.g.

— ہاں کسی ایک خرابی کی وجہ سے آپ اسے برا نہیں کہہ سکتے ، بہت اچھا موبائل ہے یہ

میرے پاس تقریباً ایک سال سے یہ فون ہے اور یہ جیسا کہ اس دن تھا جب خریدا تھا ویسا ہی اب بھی اچھا چلتا ہے -

#### **6.5.4 Testing**



All the units developed in the implementation phase are integrated into a system after testing of each unit separately. Post integration the entire system is tested for any faults/errors and failures i.e. both individual components and the integrated whole are methodically verified to ensure that they are error-free and fully meet the requirements, then the whole system behaves according to the specifications.

In this phase, testing whole software is divided into two parts: hardware and software, and type of testing are of two types inside test and outside test.

After successfully testing the software, it is delivered to the customer for use.

#### **6.5.5 Deployment of System**

Once the functional and non functional testing is done, the product is deployed in the customer environment or released into the market.

The software is now applied by the customer to his/her own system. What the customer needs to take care of is his own system complying with the minimum system requirements of the software. He also needs to take care of any system configurations and reconfigurations on his side of the deal. Once the software is properly installed, he will begin communication with the dealers on a need-to-know basis, and help report any bugs or errors that occur.

#### **6.5.6 Maintenance**

There are some issues which come up in the client environment. To fix those issues patches are released. Also to enhance the product, some better versions are released. Maintenance is done to deliver these changes in the customer environment.

All these phases are cascaded to each other in which progress is seen as flowing steadily downwards (like a waterfall) through the phases. The next phase is started only after the defined set of goals are achieved for previous phase and it is signed off, so the name "Waterfall Model".

This phase of the model is virtually never-ending. Generally, problems with the system (which are not found during the development cycle) come up after its practical use starts, so the issues related to the system are solved after its deployment. Not all the problems come into picture directly, but they arise from time to time and need to be solved; hence this process is referred to as maintenance, even though it's still pretty much in the testing phase.

This is the final phase of the waterfall model, in which the completed software product is handed over to the client after alpha, beta testing. After the software has been deployed on the client site, it is the duty of the software development team to undertake routine maintenance activities by visiting the client site.

If the customer suggests changes or enhancements the software process has to be followed all over again right from the first phase i.e. requirement analysis.

The usually the longest stage of the software. In this phase the software is updated to:

Meet the changing customer needs

Adapted to accommodate changes in the external environment.

Correct errors and oversights previously undetected in the testing phases.

Enhancing the efficiency of the software observe that feed back loops allow for corrections to be incorporated into the model.

## **6.6 Disadvantages of Waterfall Model**

Once an application is in the testing stage, it is very difficult to go back and change something that was not well-thought out in the concept stage.

No working software is produced until late during the life cycle.

High amounts of risk and uncertainty.

Not a good model for complex and object-oriented projects.

Poor model for long and ongoing projects.

Not suitable for the projects where requirements are at a moderate to high risk of changing.



### **6.7 Waterfall Model Application**

Every software developed is different and requires a suitable SDLC approach to be followed based on the internal and external factors. Some situations where the use of Waterfall model is most appropriate are:

Requirements are very well documented, clear and fixed.

Product definition is stable.

Technology is clearly understood and is not dynamic.

There are no ambiguous requirements.

Ample resources with required expertise are available to support the product.

The project is brief.

## **7 System Estimates and Actual Outcome**

At the start of any project, we have to predict or estimate the value of cost or time, which is known as Estimation. In this chapter, we will discuss the cost estimates of our project.

### **7.1 Historical data used for estimates**

For estimation, historical data holds importance. Historical data includes legacy data or data from past experience or reference to some authentic research etc. a lot of time, effort and money is required while gathering historical data.

Time plays a vital role during estimation. In our project, for determining estimation, we have divided our tasks into small units and noted time while performing the different tasks.

### **7.2 Estimation techniques applied and results**

Time Estimation is done through work breakdown. Initially work is broken down into small units and these small units are then used to determine the time estimate of the project.

Another way to Estimate Time is called "Similar Value". The time for each task is taken same as its previous similar value. By combining this time, we estimated the total time period of our project.

#### **7.2.1 Breaking Down**

We have used the breaking down technique to estimate time for our project to be completed.

For this, we broke down our project into several small units/modules such as Data Input Module, Database Container Module, Result module, Calculation Module, Frequency Module, Data Separator Module, and Graphical Module. We gave separate time for designing the UI.

### 7.2.2 Similar Value

Once the project is broken into small units we assign time slots to each working unit according to the similar value those units took for other project.

UNITS/MODULES	ESTIMATED TIME
Data Input Module	1 week
Database Container Module	4 weeks
Result module	Modification throughout the whole time
Calculation Module	2 weeks
Frequency Module	1 week
Data Separator Module	1 week
Graphical Module	1 week
Designing the UI	Modification throughout the whole time

Table 7.1: Estimated time of Project Completion

### 7.3 Actual Results and Deviation from Estimates

The costing or pricing of our project is in terms of hours and efforts that we have put into our project, as it does not require any external medium. Therefore the cost is not in terms of money.

Following is a deviation chart that is presented to illustrate the comparison between actual time of system and estimated time:

UNITS/MODULES	ESTIMATED TIME	ACTUAL TIME
Data Input Module	1 week	2.5 weeks
Database Container Module	4 weeks	Modification throughout the whole time
Result module	Modification throughout the whole time	Modification throughout the whole time
Calculation Module	2 weeks	3 weeks
Frequency Module	1 week	1.5 week
Data Separator Module	1 week	2 week
Graphical Module	1 week	1 week
Designing the UI	Modification throughout the whole time	Modification throughout the whole time

Table 7.2: Estimated Vs. Actual time of Project completion

## 7.4 System Resources (Required and Used)

People, hardware, software, tools, and other resources proposed to build the software are noted here. On the bases of which cost estimations were performed.

### 7.4.1 System Resources Required

People, hardware, software, tools, and other resources proposed to build the software are noted here. On the bases of which cost estimations were performed.

#### People:

Two types of people are the requirements for Sentiments Analysis on Urdu Text.

User/Customers:

These are the users who use the software product on the front-end.

Programmers:

These are the people who manage the system (just for the first time because it is a desktop application)

**Hardware:**

Any desktop computer /laptop.

**Software:**

The softwares used in the development of our project Sentiments Analysis on Urdu Text were not costly and they were easily available. These softwares include:

**Microsoft Visual Studio 2010:**

Microsoft Visual Studio 2010, a family of products, tools and technologies, is used for building powerful, versatile, high performance applications as well as for performing basic development tasks. Visual Studio 2010 comes with integrated support for test-driven development and debugging tools that help to ensure high quality solutions. This software is easily available. The cost of installing and buying VS 2010 is not too high. Sentiments Analysis on Urdu Text is purely based and developed on C# in Visual Studio 2010.

**Microsoft Access 2007:**

Microsoft Access 2007 is basically used for creating databases. We have used MS Access 2007 for creating our data dictionary, positive and negative columns

### **7.4.2 System Resources Used**

We have used the system resources according to our requirements.

## **8 Test Plan**

Chapter 8 is “Testing”, this chapter contains the test plan, test cases you developed and testing methods applied.

### **8.1 System Test and Procedure**

While testing a system, a test plan is needed because it is a systematic approach by which a system is tested. The plan contains a detailed description of the eventual workflow. The focus of this document is to describe:

What to test?

How to test?

When to test?

Who will do what test?

In software testing, a test plan gives detailed testing information regarding an upcoming testing, including Scope of testing, its Schedule, Test Deliverables, Release Criteria and Risks.

In software test planning, to make a project deliverable, you need to:

Identify the software to be tested

Identify the testing objectives

Identify test phases

Identify test approach

Identify methods and testing tools required

A test plan estimates how long it will take to complete the testing phase. There are many requirements to complete testing phases.

First, testers have to execute all test cases at least once. Afterwards, if a defect is found, you will have to fix the problem. You have to re-test the failed test case until it is functioning correctly. At the end, you need to conduct regression testing towards the end of the cycle to make sure the developers did not accidentally break any part of

the software while fixing another part. This occurs on test cases that were previously functioning properly.

Testing also includes the design layout. Design should be accurate and the layout should be presentable. Alignment of all the form controls, i.e. the text boxes, list box, buttons, etc must be accurate.

There are many different ways to perform software testing. Static testing is the testing without execution of Software. In Static Testing, software is examined manually and some Static analysis tool is used. It can start early in the life cycle. E.g.: By Verifying User Requirements.

Types of defect found in Static testing are: Missing requirements, Design defect, Syntax Error. Types of Static Testing: Review, Inspection, Walk-through.

Static Testing finds bug before you compile and it is about prevention. It is done in the done in the verification stage.

In dynamic testing, testing involves the execution of the Software. Here, the software is executed by giving set of inputs, examines its output and compared what is expected.

Types of defect found in dynamic testing are: Variables not constant, checking if output from the expected values.

Types of Dynamic Testing: Unit testing, Integration testing and System Testing.

Dynamic testing finds bug after compilation. It is done at the validation stage.

Testing makes sure that the system meets functional requirements. The basic reason of System Test is to find defects and correct them before the system or application is launched. No approach or method guarantees a system to be entirely free of defects.

Functional testing is concerned only with the functional requirements of a system or subsystem and tells how well the system performs its tasks and functions. These include any user commands, data manipulation, searches and user screens. Functional testing is done using the functional description provided by the client or by using the design specifications like use cases.

Non-functional testing is concerned with the non-functional requirements and is designed specifically to evaluate the readiness of a system according to the standards which are not covered by functional testing.

Basically, non-functional testing lets us measure and compare the results of testing the non-functional attributes of software systems, for example, by testing the application or system against the client's requirement or a performance requirement. Non-functional testing explains how well the product behaves as opposed to simply what the product does.

In our project we are testing our system, i.e. Sentiments Analysis on Urdu Text. Functional testing in our medium means how well each and every module of our system works, i.e. to check that the subsystems are performing their assigned tasks as they are supposed to be doing.

## **8.2 Testing strategy**

Testing strategy is basically the software testing approach to achieve testing objectives. It sets the standards for testing processes and activities, as well as other documents.

For our project, there is one Test Strategy and different number of Test Plans for each phase or level of testing. Test Strategy describes about some key issues of the testing process. This includes the testing objective, methods of testing new functions, total time and resources required for the project, and the testing environment.

The test strategy describes the test level to be performed. There are mainly three levels of testing. These are: unit testing, integration testing and system testing. Testing clarifies the major tasks and challenges of our project.

### **8.2.1 Unit testing**

The strategy and brief procedure for unit tested is described. This includes an indication of the components that went under unit tests.

Unit testing simply means to take the smallest piece of software that is to be tested, separate it from the remainder of the code, and determine whether it behaves exactly as you expect. Each unit is tested separately before integrating them into modules to test them altogether. Unit testing is proven to be valuable and a large percentage of defects are identified during its use.



To perform unit testing, separate your program into units. Conduct the code execution tests. Test the program units one at a time. This approach makes identification of errors easier because individually all the problems of each unit is already detected.

There are several benefits of unit testing. First of all, it finds a problem at an early stage and later on it makes sure the module still works correctly. The procedure is to write test cases for all functions and methods, so that whenever a change causes a fault, it can be quickly identified.

### **8.2.1.1 Unit testing Procedure**

Create unit test plan

Create testing data

Conduct test according to unit test plan

Review results of the test

### **8.2.1.2 Testing each Module**

As discussed in the previous chapters, we have divided our system into several different modules or subsystems, such as Data Input Module, Database Container Module, and Result module, Calculation Module, Frequency Module, Data Separator Module and Graphical Module. In unit testing, we took each one of these modules separately and performed testing upon them individually and then collectively after bringing them altogether.

### **8.2.1.3 Unit Cases of our System**

The components went under unit tests and several errors were recognized in this manner. When our system's algorithm was written down, our first test case was for simple positive, negative and neutral comments.

Subsequently, we moved forward towards our next task of polarity reversal. This was tested as a separate unit, isolated from the previous simple test case.

We worked upon the '↵' ("that") condition as a separate unit test and accomplished our desired results for most of the cases.

**'↵' ("that") condition**

Figure: 8.1: ‘That’ Condition

### Comma Condition

Figure 8.2 : Comma Condition

### 8.2.2 Integration testing

Integration testing is a logical extension of unit testing. It simply means that two units that have already been tested are combined into a component and the interface between them is tested. In a realistic scenario, many units are combined into

components, which are aggregated into larger parts of the program. Integration testing occurs after unit testing and before validation testing.

The basic idea is to test combinations of pieces and at the end expand the process to test your modules with those of other groups. Finally, all the modules making up a process are tested together. Beyond that, if the program is composed of more than one process, they should be tested in pairs rather than all at once.

Integration testing identifies problems that occur when units are combined. By using a test plan that requires you to test each unit and ensure the viability of each before combining units, you know that any errors discovered when combining units are likely related to the interface between units. This method reduces the number of possibilities to a far simpler level of analysis.

### **8.2.2.1 Order of Integration by System function**

In our case of polarity reversal, when unit testing was done after combining the cases, i.e. integration testing, errors evolved which were resolved later.

In case of ‘ $\sim$ ’ (“that”) condition, on combining the code with the previously finalized and tested piece of code; a lot of issues came forward which were generating inaccurate results. For this reason, we eradicated this code.

Similarly, we tested for the comma case separately and during integration testing, we did not find it much satisfactory to be joined with the system.

Thus, integration testing finds errors in complete functions and processes within and between units. It ensures that everything has been linked together correctly.

### **8.2.2.2 Integration Test Conditions**

One or more test conditions are prepared for integrating each program unit.

Some Conditions to keep in mind:

Are all software units included in integration testing?

Is the processing of each unit validated before integration testing?

Are all files used by the system being tested included in integration testing?

Test conditions are prepared for integrating each program unit. When a unit test fails it is very easy to understand why since the scope is very narrow. When an integration

test fails, things are not so simple. Because by definition an integration tests is based on many components and a specific data flow, identifying the failure cause is not always straightforward.

### **8.2.3 Validation testing**

Validation is the process of checking that a software system meets its needs and that it fulfills its basic purpose for which it was designed. It is also referred as software quality control. Validation is done at the end of the development process and takes place after verifications are completed.

During verification if some defects are missed then during validation process it can be caught as failures.

Our system is validated for the following:

The browse button is validated, so that a text file is selected only.

Check button is validated, to make sure that a file is first browsed before getting checked.

The Go button for frequency checking is validated such that previously the Check button must be clicked.

Generate separate result files button is validated such that previously the Check button must be clicked.

### **8.2.4 High-order testing (a.k.a. System Testing)**

The testing conducted on complete integrated system to evaluate the system's working with its specified requirements. System testing falls within the scope of black box testing that does not have to deal with the code or logic. It only deals with the functional requirements of the system.

System testing takes as input all the components that have successfully passed integration testing and also the software system itself integrated with any applicable system.

#### **8.2.4.1 Security testing**

In order to secure our data from being manipulated and the results to be altered, we keep our input Rich text box, positive/negative words rich text boxes and the result's textbox value as 'Read only'.

### 8.2.4.2 Stress testing

Stress testing is a form of intense testing that determines how stable a system can be. It involves testing beyond normal operational capacity, often to a breaking point, in order to observe the results. Reasons can include:

- to determine breaking points or safe usage limits
- to confirm that specifications are being met properly
- to determine modes of failure
- to test stable operation of a part or system outside standard usage

Stress testing helps to reveal understated or minute bugs that would otherwise go undetected when the application is deployed. Stress testing is therefore a necessity to be performed at an early phase of application development because it is better to fix the small bugs at the source. [10]

## 8.3 Testing resources and staffing

No specialized testing resources are purchased, only black box testing that is based on the desired output by providing an input.

Our project group comprises of 4 members, who at each stage individually performed their assigned tasks, whether it is data collection, logic design, and implementation or testing each of the system's module.

## 8.4 Test metrics

The testing matrices used in this project are the validating of data.

For example how many defects are existed within a module?

How many test cases are executed per person?

And what is the Test coverage %?

## 8.5 Testing tools and environment

No specific testing tools have been used and no unrelated hardware is used in the development of this project. The testing environment is same as the application development environment, i.e. Microsoft Visual Studio. Functions have been tested according to the black box testing and the internal code is tested by unit testing.

This application of Sentiments analysis is tested on several different kinds of reviews/comments obtained from different sites, face book and twitter which are taken as our testing data. [11]

### **8.6 Test record keeping and test log**

The test log is used to maintain a chronological record of all tests and their results. The test logs are very essential in software testing, because the logs give you deep analysis of test results and help you quickly locate and fix errors. This is an important phase because it is necessary to develop a log file that maintains record of our used log

## **9 Future Enhancements and Recommendations**

Chapter 9 is “Future Enhancements and Recommendations”; this chapter contains description about the deviation observed from the proposed system that can be recommended as future work.

### **9.1 Proposed future work**

Sentiments Analysis doesn't provide 100% accuracy. There are always chances of errors, but it gives an approximate idea of positivity, negativity or neutrality in a given text by extracting peoples' opinions.

Sentiments Analysis on Urdu Text is basically a research-based project. This work done by our team is an example to follow, for the people who are interested in research on Sentiments Analysis on Urdu Text. In future, they can definitely make a better version of it with further improvement.

A meeting was held at the end of May 2014 with Sir Sohail Sattar and Dr. Tafseer Ahmed (DHA Suffa University, Karachi) in which it was decided that a Rule-Based Approach will be followed and coding will be done in C# (Visual Studio) because Language Tools (Weka/Rapid Miner) didn't provide Urdu Support.

In the future, if any team redevelops this project on one of these NLP tools, they can take it to an even better and modernized level.

Furthermore, it is important to resolve the spacing issue, because in Urdu, spacing does not mean that a word has ended. Even the words with two or more spaces between them are sometimes a single word.

In future, if the Zero non-joiner work is done, it will definitely improve the accuracy of the results. In this way, the word dictionary will also be corrected by itself.

Logic can be further enhanced for:

Comparison sentences, because comparison is important while noting sentiments for a specific brand.

For this, it will be necessary to detect your brand name and what sentiment is associated with it to detect whether it is positive or negative.

Examples:

- 1) (سیمسنگ برانڈ ایپل کے مقابلے میں نااہل ثابت ہے)  
(Samsung brand apple kay muqablay main na ehl sabit hai)
- 2) (سیمسنگ موبائل برانڈ ایچٹیسے سے بہتر ہے)  
(Samsung mobile brand HTC sey behtar hai)

Sentences with no word that indicates any sentiment

Examples:

- 1) (لگتا ہے کہ اسے ۲ یا ۳ بار ایک دن میں چارج کرنے کی ضرورت ہے)  
(lagta hai keh isay 2 ya 3 bar aik din main charge karnay ki zaroorat hai)
- 2) (یہ فون ضرور خریدنا چاہئے)  
(yeh phone zaroor khareedna chaahiye)

Words that act as positive/negative/neutral according to a situation

Examples:

- 1) (پہلے سیمسنگ اچھا تھا، اب بہت خراب ہو گیا ہے -)  
(pehlay Samsung acha tha, ab bohat kharab hogaya hai)  
- سیمسنگ پہلے نمبر پر ہے  
(Samsung pehlay number par hai)

The word 'پہلے' (pehlay) acts as a neutral word in the first sentence, but in the second sentence it is showing positivity.

- 2) (اس کی بیٹری بہت جلدی کم ہو جاتی ہے)  
(iski battery bohat jaldi kum hojati hai)  
یہ لوڈنگ میں بہت کم وقت لگاتا ہے



(yeh loading main bohat kum waqt lagata hai)

The word “کم”(less) acts as a negative word in the first sentence but in the second sentence, its showing positivity.

Sentences for which Sentiment detection is complex

1) (سامسنگ میں بہت بہتری آگئی ہے)

(Samsung main bohat behtri aa gayi hai)

2) (سامسنگ کو بہتری لانے کی ضرورت ہے)

(Samsung ko behtri laanay ki zaroorat hai)

Both sentences contain the positive word بہتری (behtri), but the sentence structure is such that the first sentence has a positive sentiment and the second has a negative sentiment.

It will be difficult to get correct results for some types of sentences, examples are:

1) (اللہ کا شکر ہے میرے فون نے آج تک مجھے شکایت کا موقع نہیں دیا)

(Allah ka shukar hai kay meray phone nay aaj tak mujhe shikayat ka mauka nahi dia)

2) (جلدی سامسنگ خرید لو ورنہ پچھتاؤ گے)

(jaldi Samsung khareed lo warna pachtao gey)

## 9.2 Aspect-Based Sentiments Analysis

The basic elements of sentiments analysis are Entities (e.g. Mobile phone, camera, laptop, etc) and their aspects (battery, screen, etc).

In Our project we are not dealing with the aspects. This must be done in the future, because this is the way to identify the aspects of given target entities and the sentiment expressed towards each aspect.

Suppose you detect a comment as negative, and then you would want to know which aspect of that review is negative, in order to improve your product. This is aspect based sentiments analysis.

For Example:

1) (سامسنگ کے کمیرے کی بیٹری جلدی خراب ہو جاتی ہے)


(Samsung kay cameray ki battery jaldi kharab ho jati hai)

In this sentence, Camera is the entity and Battery is its aspect, which needs improvement.

## 10 Conclusion / Summary

Sentiments Analysis doesn't provide 100% accuracy. There are always chances of errors, but it gives an approximate idea of positivity, negativity or neutrality in a given text by extracting peoples' opinions. We have collected 1000 comments on Samsung products we have manually labeled the comments as positive, negative, neutral or questions. And then we inputted the comments file in our system. There were different between our manually checked results and system generated results.

The results were as follows:



<b>1000 comments</b>	<b>TOTAL</b>	<b>DETECTED</b>	<b>ACCURACY IN %</b>
<b>POSITIVE</b>	<b>543</b>	<b>499</b>	<b>92</b>
<b>NEGATIVE</b>	<b>347</b>	<b>344</b>	<b>98</b>
<b>NEUTRAL</b>	<b>91</b>	<b>134</b>	<b>61</b>
<b>QUESTION</b>	<b>23</b>	<b>23</b>	<b>100</b>

Table 10.1: Result Table

As you can see the actual positive comments were 543 but our system detected 499 positive comments. Negative comments were 347 but our system detected 344 comments. Neutral comments were 91 but system detected them as 134 it means that it has detected wrongly some of the positive and negative comments as neutral comments. Questions were detected with 100% accuracy.

Graphical representation of the result:

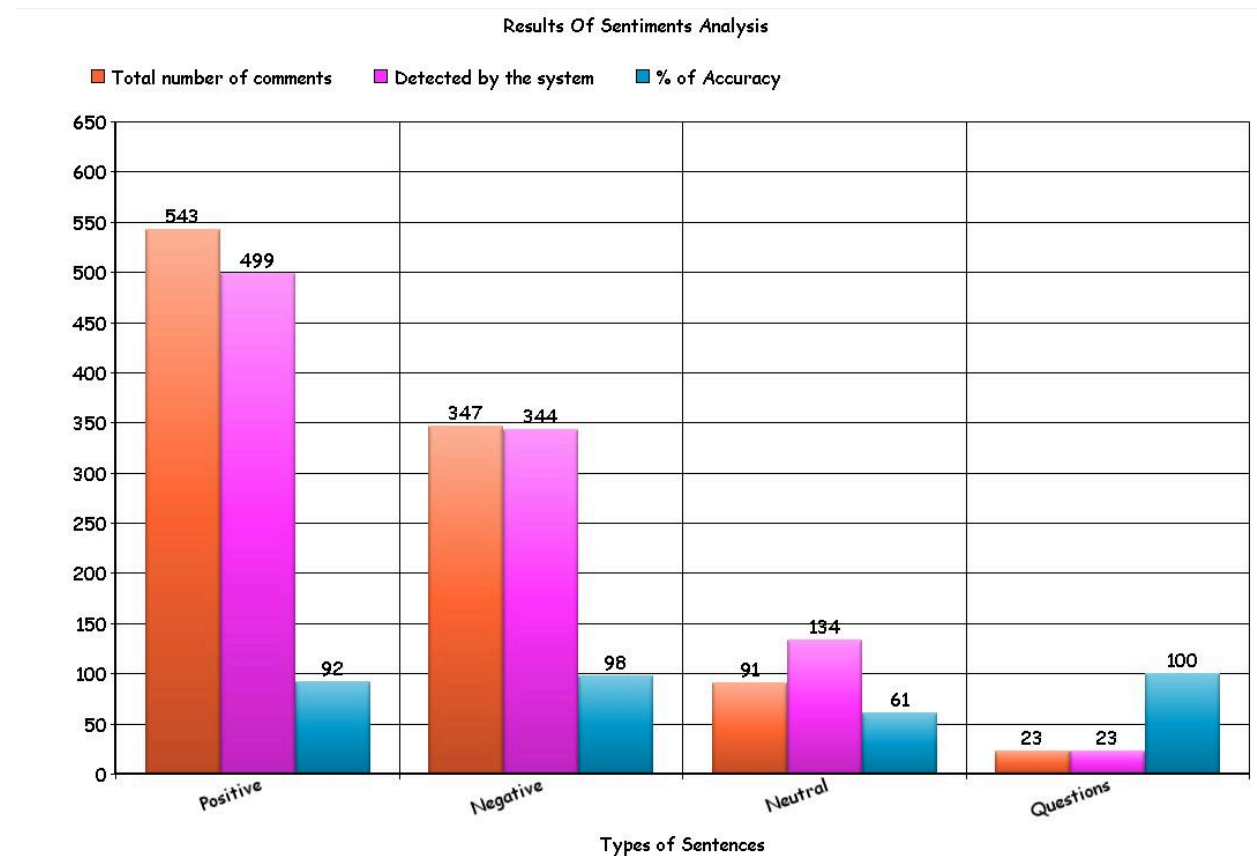


Figure 10.1: Graphical representation of Result Table

This work done by our team will be an example to follow for the people who are interested in research on Sentiments Analysis on Urdu Text in future, and they can definitely make a better version of it with further improvement.

## J. APPENDICES

### i. Project Schedule

This section presents an overview of project tasks and the output of a project scheduling tool.

**Timeline chart / Gantt Chart**

Research & planning					
Data collection, creation Of Data Dictionary and Research Paper Writing					
Algo designing and Coding					
Coding					
Coding and testing the system for results					
	January-March	April-June	July-August	September-October	November

### Project Group Organization / Work Load Distribution

The manner in which group is organized and the mechanisms for reporting are noted:

Group Members	Work Load Description
Nimra Ahmad CT-004	Data Collection Research paper Writing Data Dictionary Graphical module User interface Technical writing Testing
Qunoot Ahmed CT-021	Data Collection Research paper Writing Data Dictionary Frequency module User interface Technical writing Testing
Rabail Ali CT-040	Data Collection Research paper Writing Logic building Data separator module Calculation module Technical writing Testing
NidaSaeed CT-50	Data Collection Research paper Writing Algo writing Data input module Result module Technical writing Testing

## Working Session / Snap shots of deployed system

SENTIMENTS ANALYSIS ON URDU TEXT/ جذبات کی جانچ

Separate Result for each comment:

Browse File

اچھا  
بہترین  
شندار  
خوبصورت

بدقسمتی  
مسائل  
مایوس  
خوفناک  
خراب

Calculations

Positive:

Negative:

Neutral:

Questions:

FREQUENCY

Negative Words

Positive Words

Ascending Descending GO

Generate Separate Result Files

نتیجہ / Result

Positive  
Negative  
Neutral

جملے / Sentences

سیمسٹگ موبائل مارکیٹ میں آج بہترین فروخت ہونے والے فونز ہیں۔ بدقسمتی سے سیمسٹگ انٹرفیس کے ساتھ مسلمہ ہے کیا یہ ضروری ہے کہ اس کے بر ماڈل میں بہتری کی شکایت ہو؟ کیا اس سمارٹ ٹی وی میں کوئی خصوصیت ہے؟ کیا سیمسٹگ کا ایپ ٹاپ واقعی برا نہیں ہے؟ کیا سیمسٹگ موبائل قابل خرید ہے؟ سیمسٹگ سب سے بہتر ہے اور اس کی کارکردگی بہت اچھی ہے۔ یہ اچھی طرح سے بنیادی کالز کرتا ہے اور کال کا معیار اچھا ہے۔ جب یہ فون جاری کیا گیا تھا، اس میں بہترین صلاحیت تھیں۔ بدقسمتی سے کچھ بہتری کے مسائل ظاہر ہو گئے ہیں، بہتری کی زندگی بہت زیادہ نہیں ہے، اب کو اپنا فون چارج کرتے رہنا ہوگا۔ ان فونز کے کیمرے کے نتائج عظیم ہیں اور یہ حیرت انگیز تصاویر نکلتے ہیں۔ یہ فون کبھی ریستارٹ ہوجاتا ہے، یہ اپنے طور پر بند ہوجاتا ہے، کبھی یہ اٹکتا ہے۔ جاؤ اور اس سب کو خریدو!! اس قیمت میں یہ سب سے اچھا ہے۔ بلکا پھلکا ہے، ویڈیوز آسانی سے چلتی ہیں اور کھیل بھی چل رہے ہیں۔ کیمرے کے نتائج مقامی پرائڈز کے مقابلے میں بہت بہتر ہیں۔ یہ ایک اچھا فون ہے اور میں دسمبر سے اس کا استعمال کر رہا ہوں۔ بڑی سکرین اور خوبصورت ٹچ، اس کے لئے جاؤ! یہ ایک اچھا موقع ہے۔ اسے کھونا نہیں ہے۔ اس پخت میں ایک بہت اچھا فون ہے۔ مجھے اس کی خصوصیات سے محبت ہے۔ اینڈرائڈ اس سستی قیمت میں ناپاب ہے۔ سب سے پہلے اس فون کے ساتھ اٹکتے کا مسئلہ ہے۔

Check

منفی / Negative

مثبت / Positive

SENTIMENTS ANALYSIS ON URDU TEXT/ جذبات کی جانچ

Separate Result for each comment:

positive/مثبت  
negative/منفی  
Question  
Question  
Question  
positive/مثبت  
positive/مثبت  
negative/منفی  
positive/مثبت

Calculations

Positive: 123 49.00398%

Negative: 90 35.85658%

Neutral: 34 13.54582%

Questions: 4 1.593626%

FREQUENCY

Negative Words

Positive Words

Ascending Descending GO

Generate Separate Result Files

نتیجہ / Result

Positive

Positive  
Negative  
Neutral

جملے / Sentences

سیمسٹگ موبائل مارکیٹ میں آج بہترین فروخت ہونے والے فونز ہیں۔ بدقسمتی سے سیمسٹگ انٹرفیس کے ساتھ مسلمہ ہے کیا یہ ضروری ہے کہ اس کے بر ماڈل میں بہتری کی شکایت ہو؟ کیا اس سمارٹ ٹی وی میں کوئی خصوصیت ہے؟ کیا سیمسٹگ کا ایپ ٹاپ واقعی برا نہیں ہے؟ کیا سیمسٹگ موبائل قابل خرید ہے؟ سیمسٹگ سب سے بہتر ہے اور اس کی کارکردگی بہت اچھی ہے۔ یہ اچھی طرح سے بنیادی کالز کرتا ہے اور کال کا معیار اچھا ہے۔ جب یہ فون جاری کیا گیا تھا، اس میں بہترین صلاحیت تھیں۔ بدقسمتی سے کچھ بہتری کے مسائل ظاہر ہو گئے ہیں، بہتری کی زندگی بہت زیادہ نہیں ہے، اب کو اپنا فون چارج کرتے رہنا ہوگا۔ ان فونز کے کیمرے کے نتائج عظیم ہیں اور یہ حیرت انگیز تصاویر نکلتے ہیں۔ یہ فون کبھی ریستارٹ ہوجاتا ہے، یہ اپنے طور پر بند ہوجاتا ہے، کبھی یہ اٹکتا ہے۔ جاؤ اور اس سب کو خریدو!! اس قیمت میں یہ سب سے اچھا ہے۔ بلکا پھلکا ہے، ویڈیوز آسانی سے چلتی ہیں اور کھیل بھی چل رہے ہیں۔ کیمرے کے نتائج مقامی پرائڈز کے مقابلے میں بہت بہتر ہیں۔ یہ ایک اچھا فون ہے اور میں دسمبر سے اس کا استعمال کر رہا ہوں۔ بڑی سکرین اور خوبصورت ٹچ، اس کے لئے جاؤ! یہ ایک اچھا موقع ہے۔ اسے کھونا نہیں ہے۔ اس پخت میں ایک بہت اچھا فون ہے۔ مجھے اس کی خصوصیات سے محبت ہے۔ اینڈرائڈ اس سستی قیمت میں ناپاب ہے۔ سب سے پہلے اس فون کے ساتھ اٹکتے کا مسئلہ ہے۔

Check

بدقسمتی  
شکایت  
نہیں  
بدقسمتی  
مسائل  
نہیں  
پند

منفی / Negative

بہترین  
خصوصیت  
قابل  
بہتر  
اچھی  
اچھی  
معیار

مثبت / Positive

## **K. REFERENCE**

List and number all bibliographical references at the end of your proposal. When referenced in the text, enclose the citation number in square brackets, for example [1]. Where appropriate, include the name(s) of editors of referenced books.

- [1] Daniel Jurafsky and James H. Martin, second edition, Prentice Hall, 2009.
- [2] PrabuPalanisamy, Vineet Yadav and HarshaElchuri, “Serendio: Simple and Practical lexicon based approach to Sentiment Analysis”, Serendio Software Pvt Ltd Guindy, Chennai 600032, India, 2013.
- [3] G.Vinodhini and RM.Chandrasekaran, "Sentiment Analysis and Opinion Mining: A Survey”, International journal Of Advanced Research in Computer Science and Software Engineering, Volume 2,2012.
- [4] Afraz Z. S., A. Muhammad and Martinez-Enriquez A. M "Sentiment-Annotated Lexicon Construction for an Urdu Text Based Sentiment Analyzer", Pakistan Journal of Science, Volume 63, Pakistan, 2011.
- [5] Afraz Z. S., A. Muhammad and Martinez-Enriquez A. M “Adjectival Phrases as the Sentiment Carriers in the Urdu Text", Journal of American Science, United States, 2011.
- [6] <http://www.google.com/intl/ur/inputtools/try/>
- [7] <http://chaoticity.com/urdu-sentiment-lexicon/>
- [8] <http://raitonambele.blogspot.com/2012/06/waterfall-model.html>
- [9] <http://sfcdsrini.blogspot.com/2014/10/what-is-waterfall-model-in-sdlc.html>
- [10] [http://en.wikipedia.org/wiki/Stress\\_testing](http://en.wikipedia.org/wiki/Stress_testing)
- [11] <http://blog.design48.net/>



## **L. GLOSSARY (SMALL DEFINITIONS OF TERMINOLOGY USED IN YOUR REPORT)**

**Sentiment:** a view or opinion that is held or expressed.

**Rule-Based:** This approach uses rules to make deductions or choices. Mostly it is used for Artificial Intelligence applications and research work.

**Data Mining:** Process in which data is examined from different point of views and angles, and then summarized into useful information.

**Lexicon:** Means "of or for words". It is a dictionary of a language.

**Polarity:** The state of having two opposite or contradictory tendencies, opinions, or aspects.

**Analysis:** Detailed examination of the elements or structure.

**Modules:** A set of independent units that can be used to construct a more complex structure.

**Deployment:** Once testing is done, the application is deployed or released for the client.

**Validation:** Ensuring that data inserted into an application satisfies defined formats and other input criteria.

**Test-Cases:** A case that sets a precedent for other cases involving the same question of law.

**Accuracy:** Measured from machine evaluation and human evaluation.

