A

Project-I Report

on

SENTIMENT ANALYSIS ON SOCIAL MEDIA DATA

Submitted in Partial Fulfillment of

the Requirements for the Degree

of

Bachelor of Engineering

in

Computer Science

to

North Maharashtra University, Jalgaon

Submitted by

Sheeba F. Patel
Pranjali G. Patil
Snehal P. Patil
Swati S. Thorat

Under the Guidance of

Mr. Dinesh D. Puri

DEPARTMENT OF COMPUTER ENGINEERING

SSBT's COLLEGE OF ENGINEERING AND TECHNOLOGY,

BAMBHORI, JALGAON - 425 001 (MS)

2016 - 2017

SSBT's COLLEGE OF ENGINEERING AND TECHNOLOGY, BAMBHORI, JALGAON - 425 001 (MS)

DEPARTMENT OF COMPUTER ENGINEERING

CERTIFICATE

This is to certify that the project-i entitled *Sentiment Analysis on Social Media Data* , submitted by

Sheeba F. Patel
Pranjali G. Patil
Snehal P. Patil

Swati S. Thorat

in partial fulfillment of the degree of *Bachelor of Engineering* in *Computer Science* has been satisfactorily carried out under my guidance as per the requirement of North Maharashtra University, Jalgaon.

Date: November 10, 2016
Place: Jalgaon

Mr. Dinesh D. Puri
Guide

Prof. Dr. Girish K. Patnaik   Prof. Dr. K. S. Wani
Head   **Principal**

Acknowledgement

I thank GOD almighty for guiding us throughout the Project. I would like to thank all those who have contributed to the completion of the Project and helped me with valuable suggestions for improvement.

My sincere thanks to Principal Dr. Prof. K. S. Wani sir, SSBT's COET for having provided us facilities to complete our Project. We would like to express our gratitude and appreciation to all those who gave us the possibility to complete this report. A special thanks to Dr. Prof. G. K. Patnaik sir, Head of the Computer Department, whose help, stimulating suggestions and encouragement, helped us to coordinate our project especially in writing this report. Last but not least, many thanks go to the guide of the Project , Mr. Dinesh D. Puri whose have given his full effort in guiding the team in achieving the goal as well as his encouragement to maintain our progress in track. We thank all Staff members of our college and friends for extending their cooperation during our project report.

Above all I would like to thank our parents without whose blessings, we would not have been able to accomplish our goal.

<div align="right">

Sheeba F. Patel
Pranjali G. Patil
Snehal P. Patil
Swati S. Thorat

</div>

Contents

List of Figures

Abstract

As increasing demand of social networking sites brought a new way of expressing individuals opinion. Social networking sites have huge amount of information. The information can be seen by other user and helps to take the decision. The sentiment analysis is done by collecting the reviews of customer which are in the form of tweets. The tweets opinions are unstructured and either positive, negative or somewhat in between the two. The previous approaches used unsupervised approaches. The unsupervised approach do not contain category and there is no accurate result. The proposed approach used supervised approach. The supervised approach, Navie Bayes machine learning algorithms used label datasets for the analysis. It automatically classifies the tweets taken from social networking sites and analyse them. Its main advantage is performance i.e. precision, accuracy will increase.

Chapter 1

Introduction

Sentiment analysis is a process where the dataset consists of emotions, attitudes or assessment which takes into account the way a human thinks. To find positive, negative or a neutral aspect is a difficult task to find in a sentence. There should be very high objective in order to summarize the review. As those reviews are written in different approach which seems difficult to judge. By using sentiment analysis of review, user is satisfied or not is come to know.

In Section 1.1 Background is described. Motivation for the project is described in Section 1.2. In Section 1.3 Problem Definition of Project is highlighted. In Section 1.4 Objectives of project is described. The Organization of Report is presented in Section 1.5. Finally, in Section 1.6 Summary is presented.

1.1 Background

As there are two types of machine learning techniques which are used for the sentiment analysis. The first technique is unsupervised and another one is supervised. As previously used unsupervised learning does not consist of a category and they do not provide with the correct targets at all and therefore conduct clustering. The proposed system is using the supervised learning which is based on labeled dataset thus the labels are provided to the model during the process. These labeled dataset are trained to produce reasonable outputs when encountered during decision- making.

1.2 Motivation

As there is fast growth of using online resources, in particular social media. As user use those media for giving their review for the particular product. Those reviews are analyze by the companies for detecting their product level in the market. As they use traditional methods for the reviews which are in the form of interviews, questionnaires and surveys to

gain feedback and insight into how customers felt about their products. That methods are very time consuming and expensive. So as by using sentiment analysis there will be valuable feedback on product and services as result in better decision making for the customer along with it help company to check their level in market.

1.3 Problem Definition

World Wide Web demand is increasing so widely and there are various sites which are collecting reviews from the user. Sentiment analysis is a process where those reviews are analyzed. The dataset consists of emotions, attitudes or assessment which takes into account the way a human thinks and what they tweet. In a sentence positive, negative sentences are difficult to understand. In the proposed system the user first login into the system for giving the comments. After that the comments are stored into the database. Then analyzing of the data is done which is taken from the database. First the data is pre-processed as by removing URL, removing special symbols and conversion to lower case is done. Next, feature vector list is formed from which later feature extraction is done as the adjective is extracted from the sentence in order to the response as positive, negative or neutral. Then those response are used by the different users or company in order to take the decision.

1.4 Objective

Main objective of this major project is to provide high accuracy with following objectives:

- Develop a system which recognize sentiments of every user.

- Develop a system which calculate intensity of detected sentiment.

1.5 Organization of Report

This report is organized into chapters. While certain basic knowledge of Sentiment Analysis system is assumed, a quick review of any concept critical to understanding the functionality of Sentiment Analysis on Social Media Data is usually provided. CHAPTER 1, titled Introduction, discussed about Sentiment Analysis on Social Media Data in which background, motivation, problem definition, objective and organization of the report is described briefly.

CHAPTER 2, titled System Analysis described literature survey, proposed system of Sentiment Analysis on Social Media Data in brief.

CHAPTER 3, titled System Requirement Specification chapter highlighted hardware requirement, software requirement, functional requirement and non-functional requirements of Sentiment Analysis on Social Media Data in brief.

CHAPTER 4, titled System Design described architectural design, data flow diagram and behavioral modeling through UML diagrams for emotion detection from text system.

CHAPTER 5, titled Result and Analysis shows snapshot of proposed system.

1.6   Summary

In this chapter, Introduction to the Sentiment Analysis on Social Media Data is described. In next chapter System Analysis of project is discussed.

Chapter 2

System Analysis

In sentiment analysis, the dataset are used and analysis is done on them. The data is classified into view that is positive, negative and neutral views. The datasets are analyze by using the unigram feature extraction technique. The framework is used as where the preprocessor is applied to the raw sentences which make it more appropriate to understand. Further, the different machine learning techniques trains the dataset with feature vectors and then the semantic analysis offers a large set of synonyms and similarity which provides the polarity of the content[4]. A project which having some literature need to be done which is followed by design proposed system. When a computer-based information system is developed, systems analysis would constitute the development of a feasibility study, involving determining whether a project is economically, socially, technologically and organizationally feasible.

In Section 2.1, Literature Survey done for the proposed solution is described. Proposed System of project is described in Section 2.2. In Section 2.3 Feasibility Study of Project is described. Finally the last Section 2.4 is of Summary.

2.1 Literature Survey

Sentiment Classification techniques are divided into machine learning approach, lexicon based approach and hybrid approach. The Machine Learning Approach applies the famous machine learning algorithms. The Lexicon-based Approach divided into dictionary-based approach and corpus-based approach which use statistical or semantic methods to find sentiment polarity. The hybrid Approach is the combination of both approaches and is very common with sentiment lexicons playing a key role in the majority of methods.The different techniques are shown in Figure 2.1.
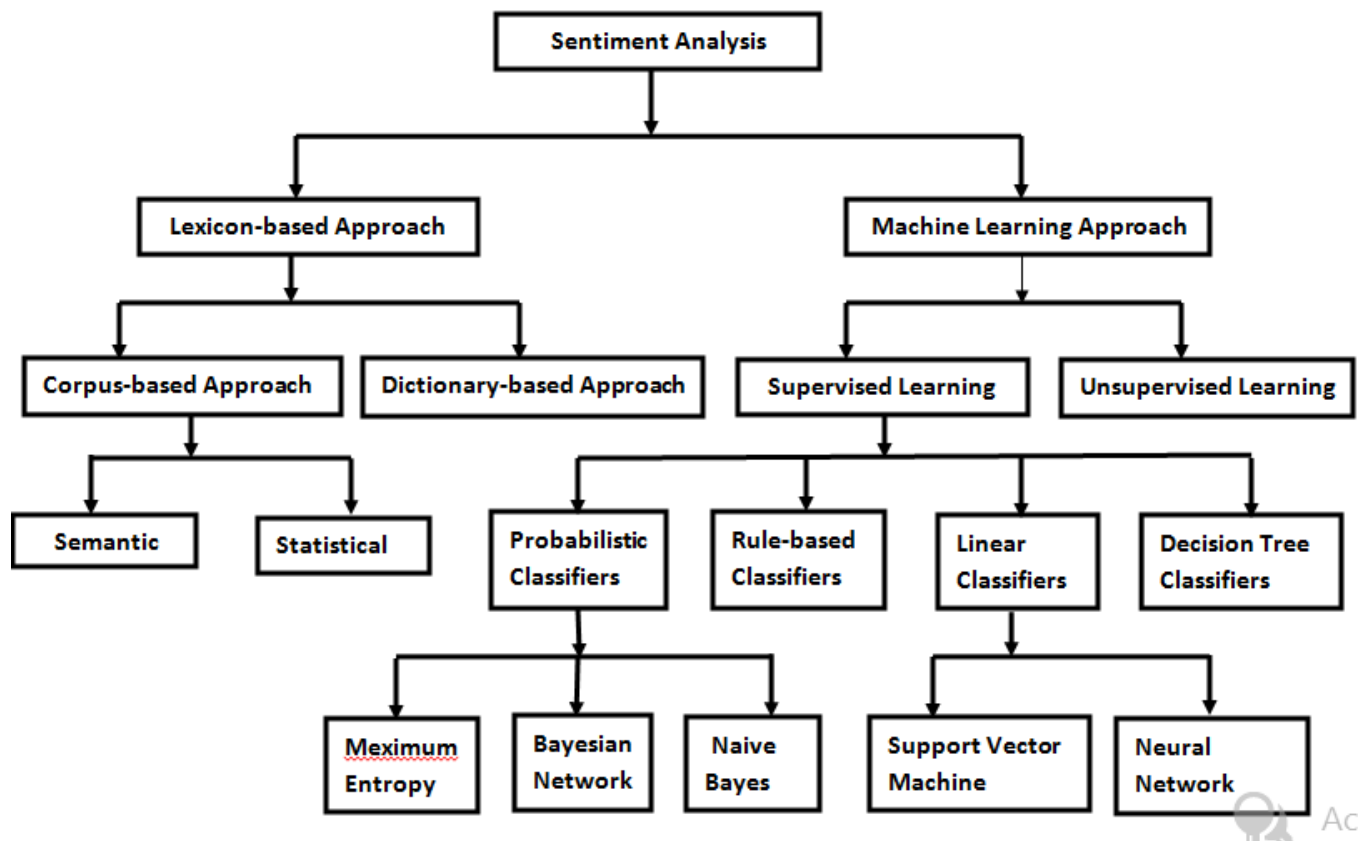
Figure 2.1: Techniques Use for semantic analyis

Further machine learning is divided into unsupervised learning and supervised learning. Then supervised learning is divided into decision tree classifiers, linear classifiers, rule-based classifiers and probabilistic classifiers. Linear classifiers are divided into support vector machines and neural network. Probabilistic classifiers are sub parts into naive bayes, Bayesian network and maximum entropy. Were lexicon-based approach are devided into dictionary based approach and the corpus based approach. The corpus based approach is sub part into statistical and semantic.

2.1.1 Lexicon Based Approach

The lexicon based approach is the approach in which opinions lexicon are find out which later used for the analyzing the text. Lexicon based approach is divided into two methods:

- **Dictionary Based Approach**, Hu Minging et. al., in[4] proposed this approach in which opinion seed words are find out and the searching is done in the dictionary of their synonyms and antonyms.

- **Corpus-Based Approach**, the Corpus-based approach helps to solve the problem of finding opinion words with context specific orientations. Its methods depend on

syntactic patterns or patterns that occur together along with a seed list of opinion words to find other opinion words in a large corpus. One of these methods were represented by Hatzivassiloglou et al., in[3]. Further divided into:

– **Statistical Approach**, the approach is given by Fahrni et al.,in[2], in which there is finding of co-currence pattern or seed opinion words can be done using this technique.

– **Semantic Approach**, the Semantic approach gives sentiment values directly and relies on different principles for computing the similarity between words. The Semantic approach is used in many applications to build a lexicon model for the description of verbs, nouns and adjectives to be used in SA as the work presented by Maks and Vossen in[8].

2.1.2 Machine Learning Algorithm

Machine learning approach relies on the famous ML algorithms to solve the Sementic Approach as a regular text classification problem that makes use of syntactic features. Divided into two methods:

- **Supervised Learning**,the supervised learning methods depend on the existence of labeled training documents, which is divided further into:

  – **Probabilistic classifiers**, Probabilistic classifiers use mixture models for classification. The mixture model assumes that each class is a component of the mixture, which further have subparts as:

    ∗ **Naive Bayes Classifier**, Proposed by Kang et al., in[6] which is most commonly and simple used classifier. Naive Bayes Classification computes the posterior probability of a class, based on the distribution of the words in the document .

    ∗ **Bayesian Network**, BN was used by Hernandez et al.,in[9] to consider a real-world problem in which the attitude of the author is characterized by three different (but related) target variables. They proposed the use of multi-dimensional Bayesian network classifiers[9].

    ∗ **Maximum Entropy Classifier**, The Maxent Classifier (known as a conditional exponential classifier) converts labeled feature sets to vectors using encoding. This encoded vector is then used to calculate weights for each feature that can then be combined to determine the most likely label for a feature set[11].

– **Linear classifiers**, this classifier is divided into 2 methods:

* **Support Vector Machines Classifiers**,given by Chen et al.,in[1]. Support vector machine analyzes the data, define the decision boundaries and uses the kernels for computation which are performed in input space. The input data are two sets of vectors of size m each. Then every data represented as a vector is classified in a particular class. Now the task is to find a margin between two classes that is far from any document. The distance defines the margin of the classifier, maximizing the margin reduces indecisive decisions.

* **Neural Network**, Neural Network consists of many neurons where the neuron is its basic unit. Multilayer neural networks are used for non-linear boundaries[10]. These multiple layers are used to induce multiple piecewise linear boundaries, which are used to approximate enclosed regions belonging to a particular class. The outputs of the neurons in the earlier layers feed into the neurons in the later layers.

- **Decision tree classifiers**, Quinlan JR, in [5] has proposed Decision tree classifier, which provides a hierarchical decomposition of the training data space in which a condition on the attribute value is used to divide the data.

- **Rule-Based classifiers**, In rule based classifiers, by using set of rules the data space is get modeled . The left hand side represents a condition on the feature set expressed in disjunctive normal form while the right hand side is the class label. The conditions are on the term presence[7]. Term absence is rarely used because it is not informative in sparse data.

2.2   Proposed System

In the given proposed system is data from the social media is taken and analysis is done on it. For analyzing those data the unigram extraction technique is used. For this first preprocessed the dataset, after that extracted the adjective from the dataset that have some meaning which is called feature vector, then selected the feature vector list and thereafter applied machine learning based classification algorithms namely: Naive Bayes with the Semantic Orientation based WordNet which extracts synonyms and similarity for the content feature which provides the polarity to the contain.

2.3 Feasibility Study

The feasibility study is an evaluation and analysis of the potential of a proposed project which is based on extensive investigation and research to support the process of decision making.Feasibility studies aim to objectively and rationally uncover the strengths and weaknesses of an existing business or proposed venture, opportunities and threats present in the environment, the resources required to carry through, and ultimately the prospects for success.

2.3.1 Economical Feasibility

Feasibility studies are crucial during the early development of any project and form a vital component in the business development process. Accounting and Advisory feasibility studies enable organizations to assess the viability, cost and benefits of projects before financial resources are allocated. They also provide independent project assessment and enhance project credibility.

2.3.2 Operational Feasibility

In its simplest terms, the two criteria to judge feasibility are cost required and value to be attained The objective of the proposed approach is to utilize the strengths of supervised approach to complement the weaknesses of other unsupervised approach techniques. As unsupervised learning does not consist of a category and they do not provide with the correct targets at all and therefore conduct clustering were as the supervised learning is based on labeled datasets and thus the labels are provided to the model during the process. These labeled datasets are trained to produce reasonable outputs when encountered during decision- making. As the supervised approach is used in proposed system as for the feasibility and effectiveness of the approach.

2.3.3 Technical Feasibility

Technical feasibility is one of the first studies, it must be conducted after the project has been identified.Technical Feasibility Study assesses the details of how you will deliver a product or service (i.e. materials, labor, transportation, where your business will be located, technology needed, etc.). Think of the technical feasibility study as the logistical or tactical plan of how your business will produce, store, deliver, and track its products or services. A technical feasibility study is an excellent tool for trouble-shooting and long-term planning. In some regards it serves as a flow chart of how the products and services evolve and move through your business to physically reach your market.

For this project the technical feasibility was studied in two aspects. One is Hardware feasibility and other is software feasibility. The system is very easy to upgrade and maintenance is also very easy. System is developed in Java and analysis in done in R language.

2.4   Summary

In this chapter, System Analysis for Sentiment Analysis on Social Media Data is described. In next chapter System Requirement Specification of project is discussed.

Chapter 3

System Requirement Specification

System Requirement Specification chapter contains the information related to what hardware and software are to be require for build system. Understanding user requirements is an integral part of information systems design and is critical to the success of interactive systems. It is now widely understood that successful systems and products begin with an understanding of the needs and requirements of the users. Requirements engineering provides the appropriate mechanism for understanding what the customer wants, analyzing need, assessing feasibility, negotiating a reasonable solution, specifying the solution unambiguously, validating the specification, and managing the requirements as they are transformed into an operational system.

In Section 3.1, Hardware Requirements are described. Software Requirement are highlighted in Section 3.2. In Section 3.3, Functional Requirements of project are described. Non Functional Requirements of project are described in Section 3.4. In section 3.5 other requirements are described. Finally Summary is presented in Section 3.5.

3.1 Hardware Requirements

In the hardware requirement, processor having minimum Intel core i3 or as system configuration which contain may be core i3. The system memory having minimum 2 to 4 GB RAM upto depend upon other system configuration. Operating system may having LINUX for software which are need to development of project.

3.2 Software Requirements

The software requirements analysis process covers the complex task of eliciting and documenting the requirements of all these users, modeling and analyzing these requirements and documenting of the basis for system design. Development of required software consist R language in which RStudio technology use as a front end. Java language in which jdk 1.6.0

technology use as a backend end and use jre, jar as application libraries.

3.3 Functional Requirements

Functional requirements define the task or function the proposed solution must do. Functional requirement defies a function of system and its components. A functions described in proposed approach is as set of inputs taken from the user as a comments, then sentiment analysis process is applied as preprocessing the data, feature extraction and getting output as the polarity to that particular comment.

3.4 Non Functional Requirements

In Non functional Requirements of this project implements those functions which does not effect on function and behavior of project for desired goal and objective of project. Non functional Requirement just provides user friendliness. The non-functional requirements of the proposed system include the basic input given by the user in text format and output can be get in the form form polarity.

3.5 Other Requirements

For developing the project other requirements can be Umbrello which is analysis and designing tool.

3.6 Summary

In this chapter, System Requirement Specification of the project is discussed. In the next chapter System Design of project is described.

Chapter 4

System Design

In this chapter, System design provides the understanding and procedural details necessary for implementing the system. In this chapter, the overall design of the proposed system is given. System design provides the understanding and procedural details necessary for implementing the system recommended in the system study.

In Section 4.1 shows Architecture for proposed system. Behavioral model of the system is described in Section 4.2.

4.1 Architecture

The Figure 4.1 shows the architecture of the proposed system.
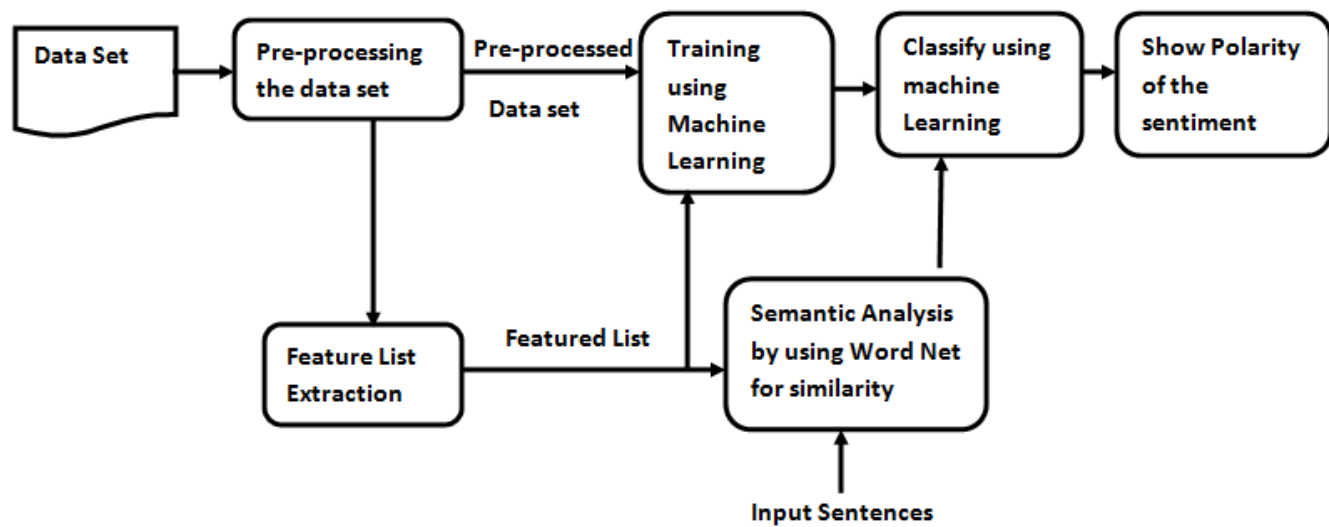

Figure 4.1: Architecture of Proposed System

The overall architecture of the proposed system can be given as:

- Pre-processing of the dataset

  As it is not easy to analyse the comments, so to make it easy to analyse pre-processing is done on those data. In pre-processing, the repeated words and the punctuation get removed to get quality of data.

- Feature Extraction

  After pre-processing the data, feature attraction take place were adjective from the datasets are get extracted. Later this adjective is used to show the positive and negative polarity in a sentence which is useful for determining the opinion of the individuals using unigram model. Unigram model extracts the adjective and segregates it. It discards the preceding and successive word occurring with the adjective in the sentences. For above example, i.e. painting Beautiful through unigram model, only Beautiful is extracted from the sentence.

- Training and classification

  As for the classification of the data some technique is to be used, proposed system is using the supervised approach of machine learning algorithm i.e. Naive Bayes approach.

4.2   UML Diagram

UML, as the name shows, is a modelling language. It is used to specify, draw, visualize and document the parts of the software. It provides a set of notations to create the visual model of a system. This phase is used to design different UML diagrams corresponding to the application development.

4.2.1   Use Case Diagram

Use case diagrams are the diagrammatic representation depicting users interactions with the system. Figure 4.2, shows use case diagram of Sentiment Analysisn on Social Media Data. The two main actors involved with usecases are user and the system. Use-case and system togetherly has 2 use-cases associated with it i.e viewing the comments and the polarity of data given by the system which will going to seen by user. In addition to this user has another usecase i.e registering and the login. Whereas the system has the use cases as removing url, removing special symbols, convert word to lower case,get feature vector list,extract features and the checking alternative synonyms.
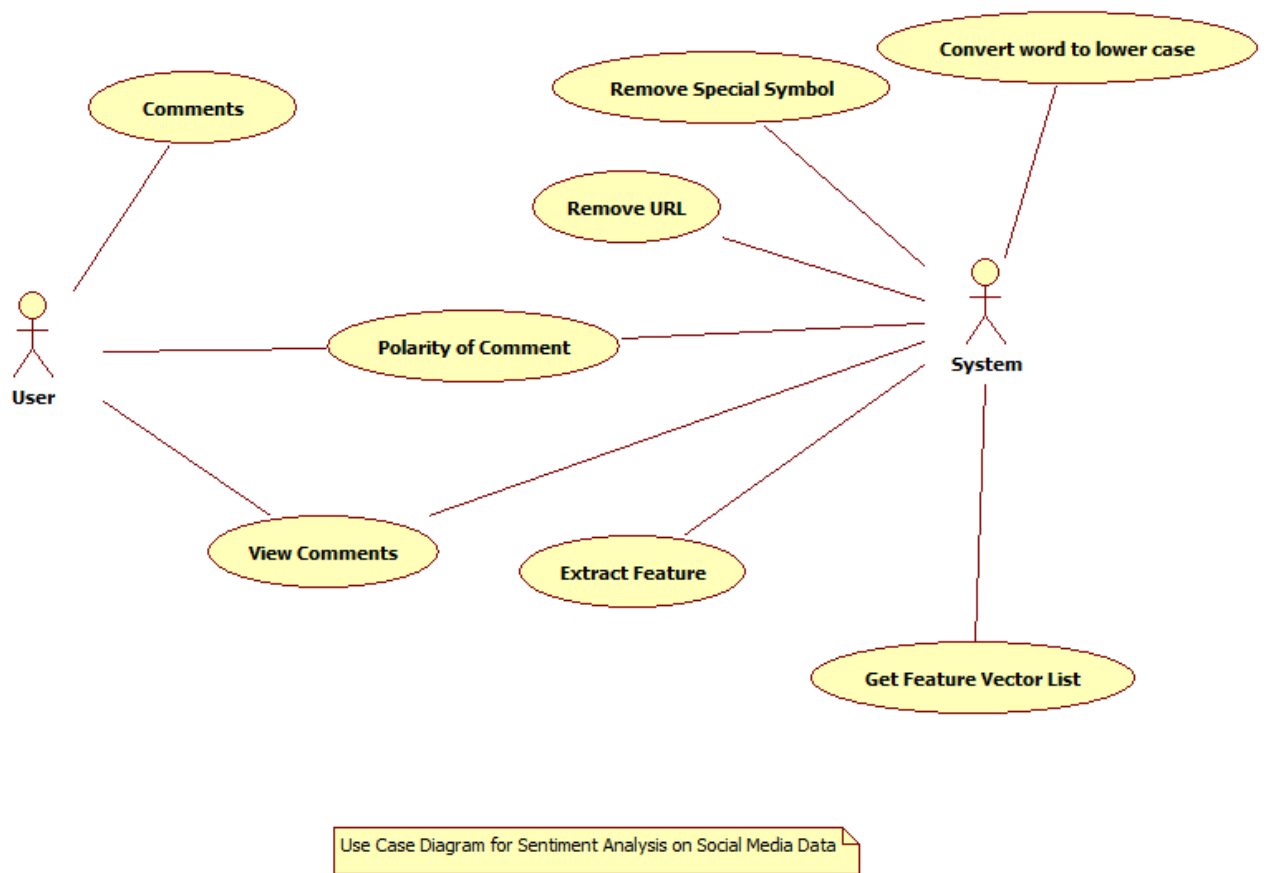
Figure 4.2: Use Case Diagram for Sentiment Analysisn on Social Media Data

4.2.2 Sequence Diagram

Sequence Diagram depicted in Figure 4.3 which has object: user, system and database. Diagram elaborates over here various level of of sentiment analysis of data through which it go through.
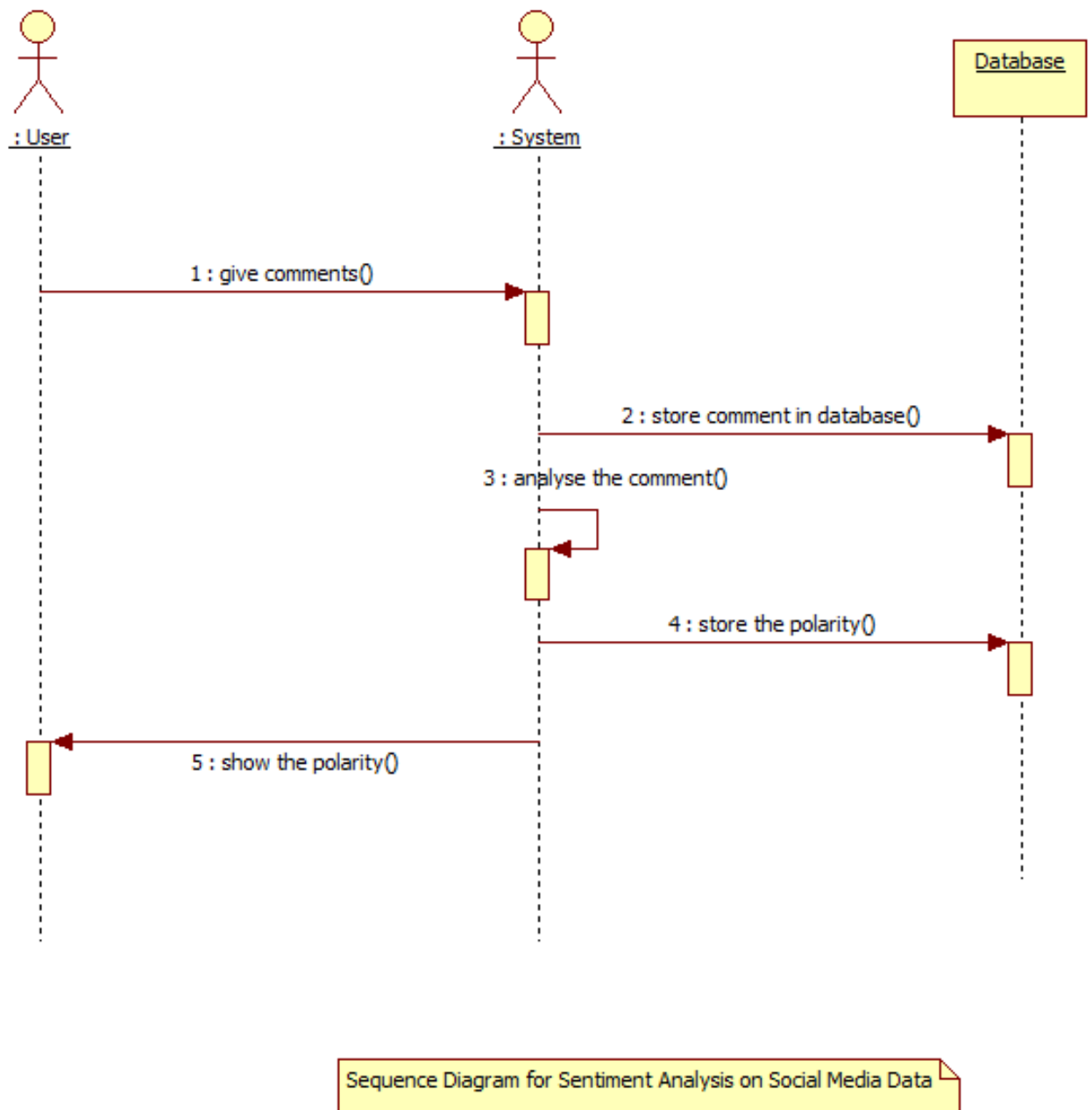
Figure 4.3: Sequence Diagram for Sentiment Analysisn on Social Media Data

4.2.3 State Transition Diagram

A state transition diagram shows a state machine, consisting of state,transitions,events and activities. State chart diagrams address the dynamic view of a system. They are especially important in modeling the behavior of an interface, class or collaboration and emphasize the event ordered behavior of an object, which is specially useful in modeling reactive system.Figure 4.4 shows the state transition diagram for the proposed system.
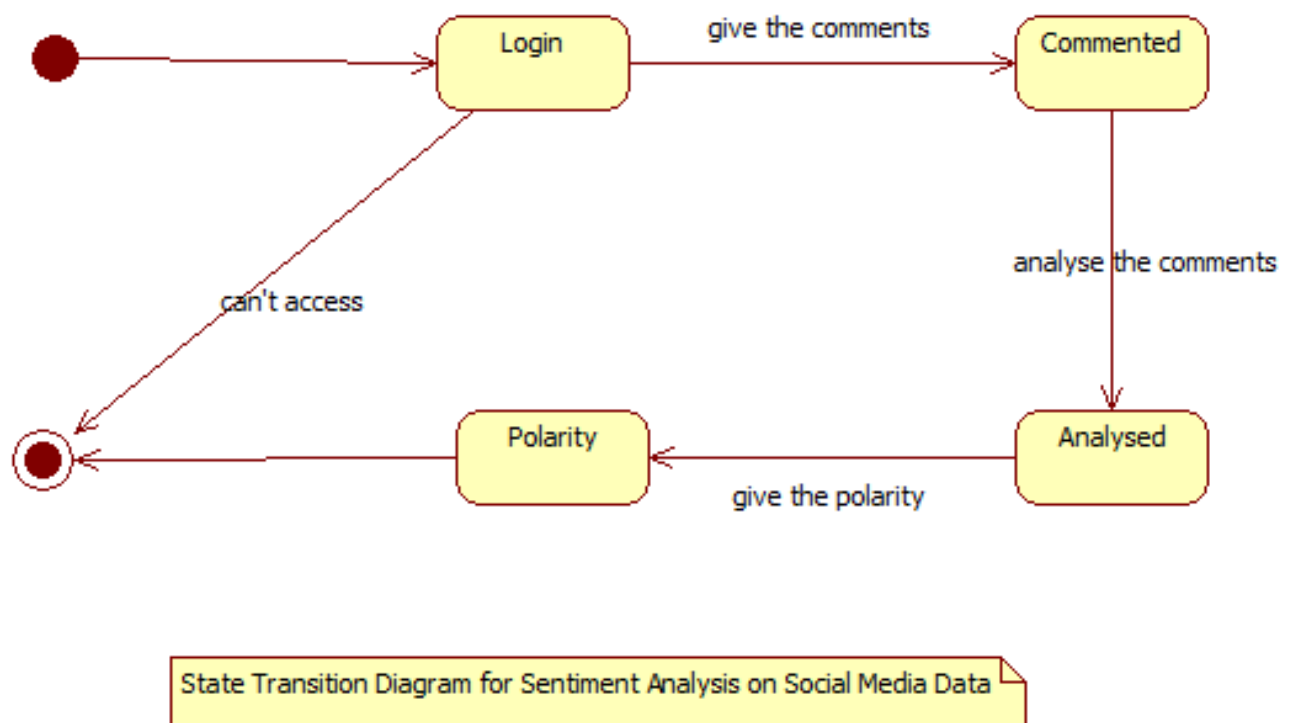
Figure 4.4: State Transition Diagram for Sentiment Analysisn on Social Media Data

4.2.4   Activity Diagram

Activity diagram is basically a flow chart to represent the flow form one activity to another activity. The activity can be described as an operation of the system.So the control flow is drawn from one operation to another.Activity Diagram shown in Figure 4.5 depicts overall activities carried in system.
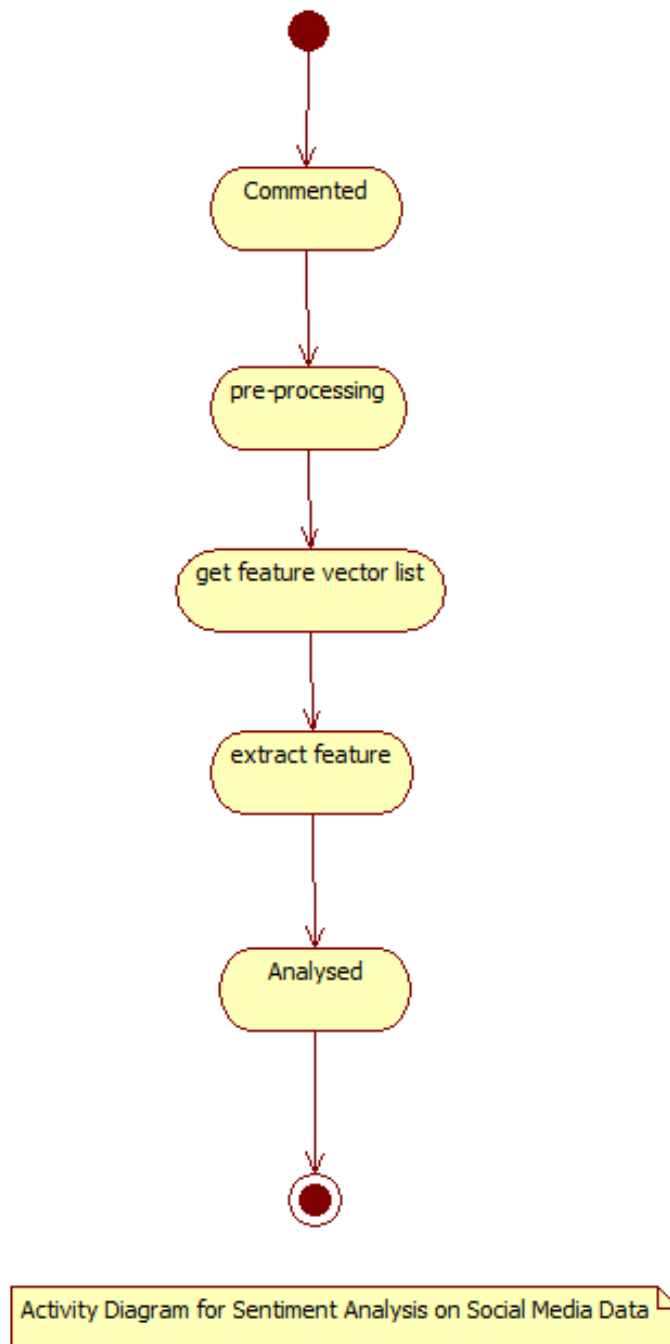
Figure 4.5: Activity Diagram for Sentiment Analysisn on Social Media Data

4.2.5 Class Diagram

Class diagram is a type of static structure diagram which describes the structure of a system by representing the classes of the system, their attributes, operations and the relationships among these classes. In the Figure 4.6, main classes defined in system are user, System Analyser, url remover, symbol remover, convertor, vector list, extractor classes.
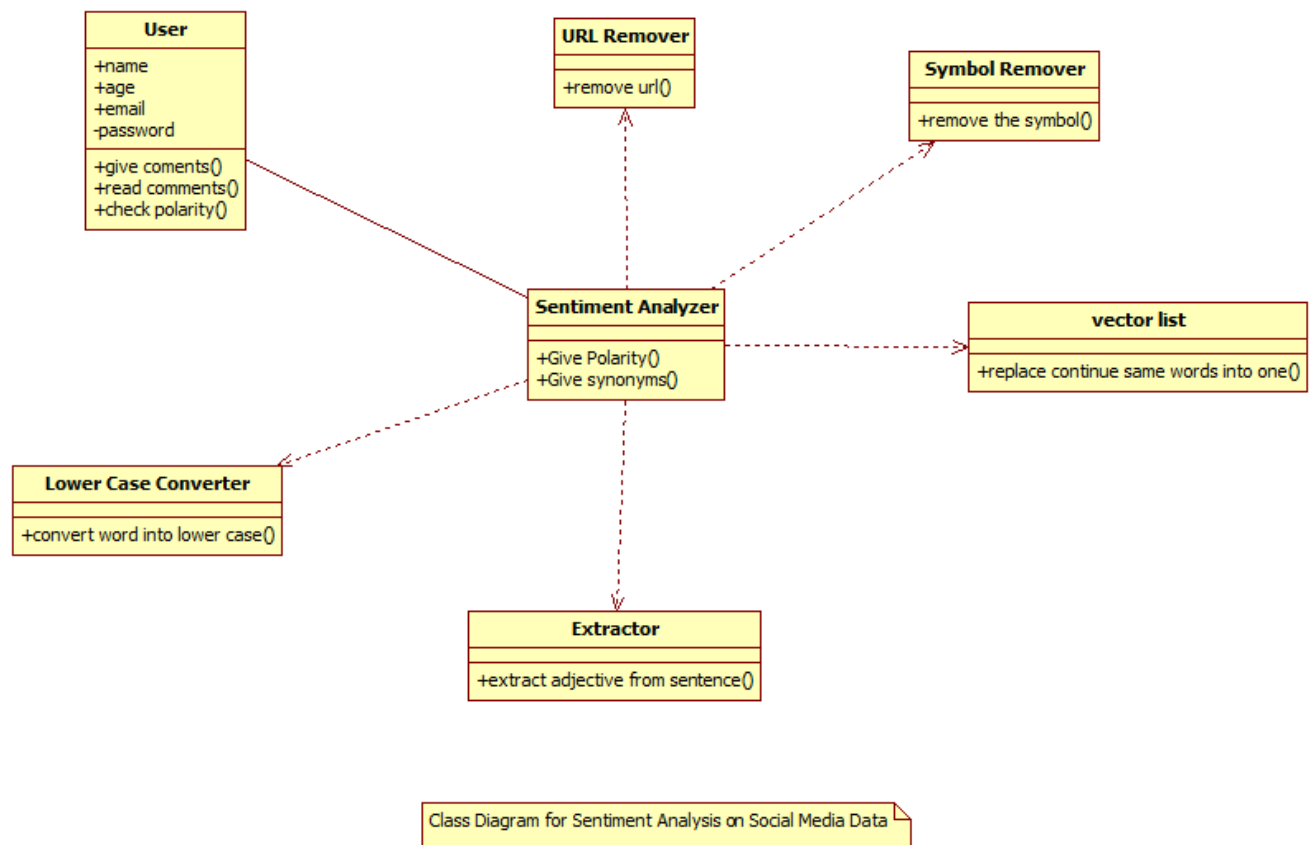
User

+name
+age
+email
-password

+give coments()
+read comments()
+check polarity()

URL Remover

+remove url()

Symbol Remover

+remove the symbol()

Sentiment Analyzer

+Give Polarity()
+Give synonyms()

vector list

+replace continue same words into one()

Lower Case Converter

+convert word into lower case()

Extractor

+extract adjective from sentence()

Class Diagram for Sentiment Analysis on Social Media Data

Figure 4.6: Class Diagram for Sentiment Analysisn on Social Media Data

4.2.6 Component Diagram

A component diagram shows the organizations and dependencies among a set of components. Component diagrams address the static implementation view of a system. They are related to class diagrams in that a component typically maps to one or more classes, interfaces, or collaborations. Component Diagram Contains Package Specification, Task specification, Object, and dependency relationship among the object and the packages. As for proposed system the main executable component is Sentiment analysis.exe. were it is depend among different component as preprocessor, feature vector list, feature extractor and classifier. Component Diagram for Sentiment Analysis on Social Media Data is shown in Figure 4.7.
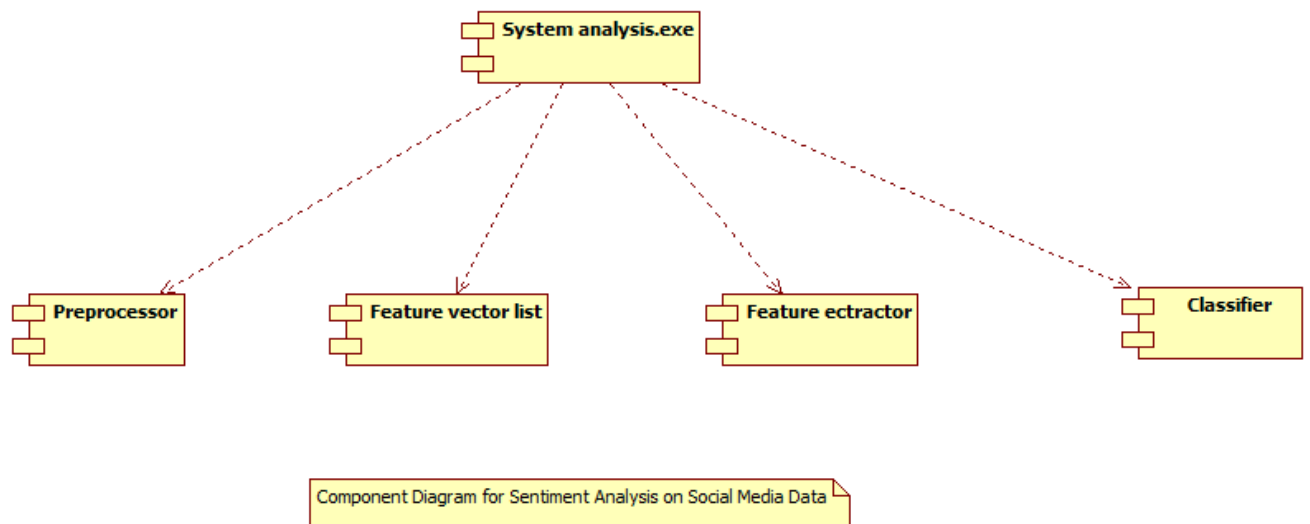
Figure 4.7: Component Diagram for Sentiment Analysisn on Social Media Data

4.2.7 Deployment Diagram

A deployment diagram shows the configuration of run-time processing nodes and the components that live on them. Deployment diagrams address the static deployment view of an architecture. They are related to component diagrams in that a node typically encloses one or more components. This Diagram contains processor and device which is mainly used to present hardware and s/w required which is having some memory. In proposed system nodes can be taken as sentiment analysis system, database and the number of users. The deployment diagram of sentiment anaylsis on social media Data is showm in Figure 4.8.
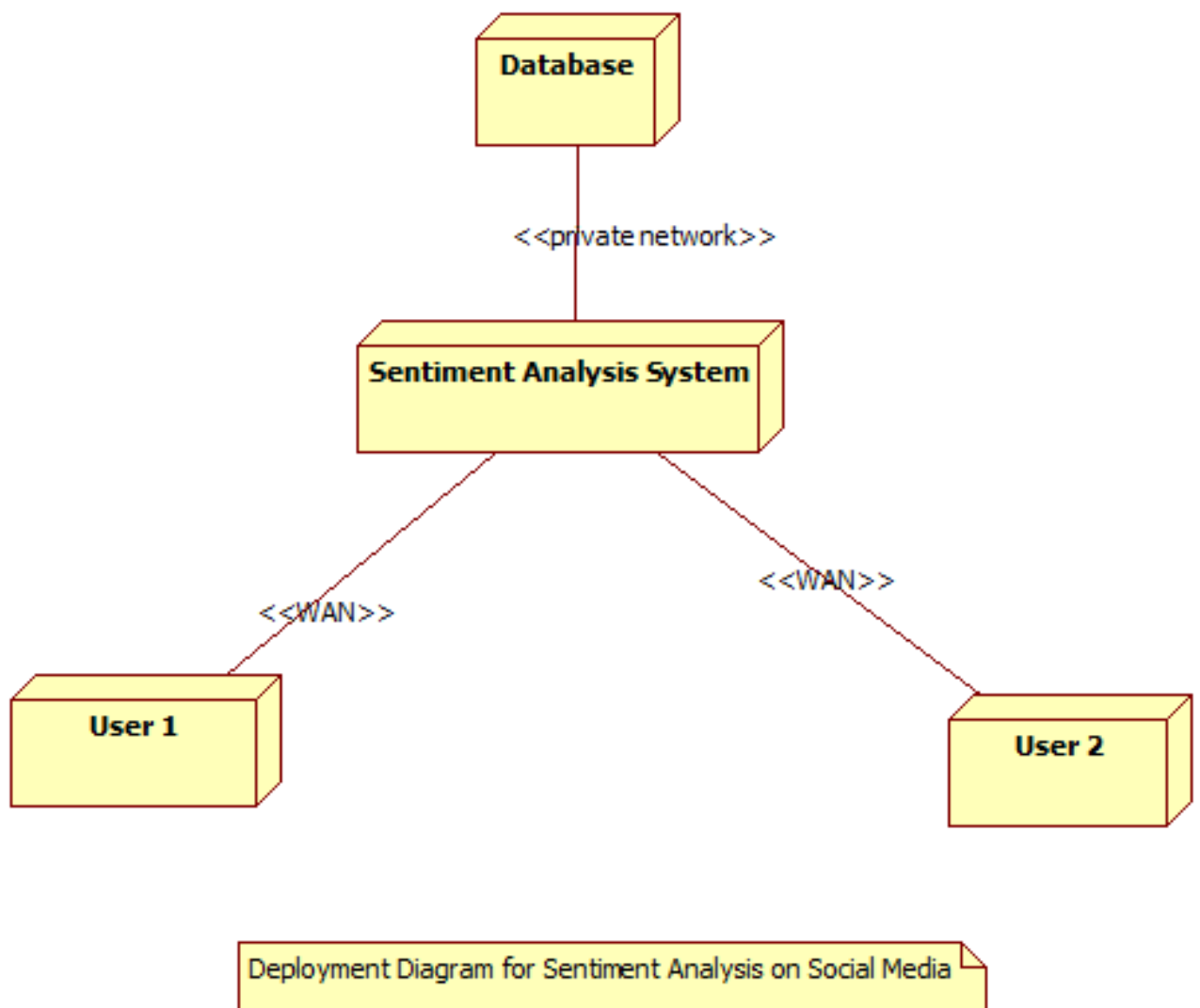
Figure 4.8: Deployment Diagram for Sentiment Analysisn on Social Media Data

4.3 Summary

In this chapter System Design is studied. In the next chapter the Expected Result along with Conclusion of the system is described.

Chapter 5

Expected Result and Conclusion

As the analysis and design of the proposed system is done. Therefore it is concluded that implementation of the proposed system can be done.

The Polarity can be get for the comments given by the user in positive, negative or in neutral form can be the expected result of the proposed system.

Bibliography

[1] Tseng You-De Chin Chen Chien. Quality evaluation of product reviews using an information quality framework. In *Decision Support System*, 2011.

[2] Klenner M Fahrni A. Old wine or warm beer: target-specific sentiment analysis of adjectives. In *Proceedings of the symposium on affective language in human and machine*, 2008.

[3] McKeown K Hatzivassiloglou V. Predicting the semantic orientation of adjectives. In *Proceedings of annual meeting of the Association for Computational Linguistics*, 1997.

[4] Liu Bing Hu Minging. Mining and summarizing customer reviews. In *Proceedings of ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, 2004.

[5] Quinlan JR. Induction of decision trees. *Machine Learn*, 1:81–106, 1986.

[6] Han Dongil Kang Hanhoon, Yoo Seong Joon. Senti-lexicon and improved naie bayes algorithms for sentiment analysis of restaurant reviews. 2012.

[7] Ma Yiming Liu Bing, Hsu Wynne. Integrating classification and association rule mining. 1998.

[8] Vossen Piek Maks Isa. A lexicon model for deep sentiment analysis and opinion mining applications. In *Decision Support System*, 2012.

[9] Rodrguez Juan Diego Ortigosa-Herna ndez Jonathan. Approaching sentiment analysis by using semi-supervised learning of multi-dimensional classifiers. In *Neurocomputing*, 2012.

[10] van den Bosch Antal van de Camp Matje. The socialist network. *Decision Support System*, 2012.

[11] Chang Pei-Chann Yu Liang-Chih, Wu Jheng-Long. Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news. In *Knowledge-Based System*, 2013.