Project Report on

HANDLING POLARITY SHIFT FOR OPINION MINING

Submitted in partial fulfillment of the requirements

of the degree of Bachelor in Engineering

by

Tanmay Shukla	B.E. - 5 - 63
Vidhan Thakur	B.E. - 5 - 73
Kandaswamy Thevar	B.E. - 5 - 74
Deepak Vishwakarma	B.E. - 5 - 77

Under the guidance of

Ms. Kranti Ghag

DEPARTMENT OF INFORMATION TECHNOLOGY

SHAH AND ANCHOR KUTCHHI ENGINEERING COLLEGE

CHEMBUR, MUMBAI- 400088.

2016-2017

www.shahandanchor.com Tel: 022 2558 0854

Mahavir Education Trust's

SHAH & ANCHOR KUTCHHI ENGINEERING COLLEGE

Mahavir Education Trust Chowk, W.T. Patil Marg, Chembur, Mumbai 400 088

Affiliated to University of Mumbai, Approved by D.T.E. & A.I.C.T.E.

Awarded provisional accreditation for Computer & Electronics Engineering by NBA

(for 2 years from 06-08-2014)

ISO 9001:2008 Certified

Certificate

This is to certify that the report of the project entitled

HANDLING POLARITY SHIFT FOR OPINION MINING

is a bonafide work of

Tanmay Shukla	B.E. - 5 - 63
Vidhan Thakur	B.E. - 5 - 73
Kandaswamy Thevar	B.E. - 5 - 74
Deepak Vishwakarma	B.E. - 5 - 77

submitted to the

UNIVERSITY OF MUMBAI

during semester VII in partial fulfilment of the requirement for the award of the degree of

BACHELOR OF ENGINEERING

in

INFORMATION TECHNOLOGY

(Ms. Kranti Ghag)
Guide

(Ms. Swati Deshpande)
I/c Head of Department

(Dr. Bhavesh Patel)
Principal

Approval for Project Report for B. E. semester VII

This project report entitled *HANDLING POLARITY SHIFT FOR OPINION MINING* by *Tanmay Shukla, Vidhan Thakur, Kandaswamy Thevar and Deepak Vishwakarma* is approved for semester VII in partial fulfilment of the requirement for the award of the degree of Bachelor of Engineering.

Examiners

1._____

2._____

Guide

1._____

2._____

Date:

Place:

Declaration

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

Tanmay Shukla	B.E. - 5 - 63	----------
Vidhan Thakur	B.E. - 5 - 73	----------
Kandaswamy Thevar	B.E. - 5 - 74	----------
Deepak Vishwakarma	B.E. - 5 - 77	----------

Date:

Date: 28/09/2016

To,
The Principal
Shah and Anchor Kutchhi Engineering College,
Chembur, Mumbai-88

Subject: Confirmation of Attendance

Respected Sir,

This is to certify that Final year (BE) students

Tanmay Shirish Shukla
Vidhan Rajnikant Thakur
Kandaswamy Sankaran Thevar
Deepak Bhoju Vishwakarma
Have duly attended the sessions on the day allotted to them during the period from 13-07-2016 to 28-09-2016 for performing the Project titled Handling Polarity Shift for Opinion Mining.
They were punctual and regular in their attendance. Following is the detailed record of the student's attendance.

Attendance Record:

Date	Tanmay Shukla	Vidhan Thakur	Kandaswamy Thevar	Deepak Vishwakarma
13-07-2016	Present	Absent	Present	Absent
20-07-2016	Present	Present	Present	Present
27-07-2016	Present	Present	Present	Present
03-08-2016	Present	Present	Present	Present
10-08-2016	Present	Present	Present	Present
24-08-2016	Present	Present	Present	Present
31-08-2016	Present	Present	Present	Present
14-09-2016	Present	Present	Present	Present
21-09-2016	Present	Present	Present	Present
28-09-2016	Present	Present	Present	Present

Ms. Kranti Ghag

Abstract

Opinion mining or sentiment analysis is the study of people's opinion, sentiments, attitudes and emotions expressed in written language. It is one of the most active research programs in natural language processing and text mining. In recent years it has gained a lot of popularity. Though sentiment analysis has been a trending research program in the community of natural language processing, the Bag Of Words based machine learning approach is state-of –the-art for this task. However, BOW model does not focus much on polarity shift which may create a different overall impact. Polarity shift handling is one of the major problems in performing sentimental analysis of any text or sentence. This is the phenomenon in which polarity of any sentence shifts from positive to negative or vice versa due to some semantics like explicit negation, explicit contrast and sentimental inconsistency. Earlier work has been done on handling the polarity shifts focused on detecting polarity shift with limited scope. Some works also included training of the classifier either by reversing of the original reviews or extracting the features based on patterns. This study aims to handle polarity shift. Sentiment classification will be performed in 3 major steps i.e. Pre-processing, Tokenization and Polarity shift handling. A model is proposed to handle explicit negation with a larger scope. Unlike traditional negation modifiers, our aim is to negate the related terms even if it does not immediately follow negation modifier. Apart from the proposed modification in explicit negation handling, other polarity shift techniques such as handling explicit contrast and sentimental inconsistency are accomplished traditionally. Apart from modification and polarity shift handling, other tasks for sentimental classification such as pre-processing, tokenization will be performed in a traditional way. Proposed model will be evaluated on Kaggle's "Bag of Words meets Bag of Popcorn" which is a balanced dataset consisting of 50,000 positive and negative reviews in total. Ten-fold cross validation technique will be used for evaluation of proposed model. Proposed model will be compared and analysed with the existing state-of-art model such as PSDEE.

Table of Contents

Abstract vi
Table of Contents vii
List of figures viii
List of Tables viii
Abbreviation, notation and nomenclature ix
Chapter 1: Introduction 1
Chapter 2: Review of literature 3
Chapter 3: Report on present investigation 6
3.1 Software Requirement Specification 6
3.1.1 Introduction 6
3.1.1.1 Purpose 6
3.1.1.2 Document Conventions 6
3.1.1.3 Intended Audience 6
3.1.1.4 Project scope 7
3.1.1.5 References 7
3.1.2 Overall description 7
3.1.2.1 Product Perspective 7
3.1.2.2 Product functions 7
3.1.2.3 User classes and characteristics 8
3.1.2.4 Operating environment 9
3.1.2.5 Design and implementation constraints 9
3.1.2.6 User documentation 9
3.1.2.7 Assumptions and dependencies 9
3.1.3 External Interface Requirements 9
3.1.3.1 User interfaces 9
3.1.3.2 Hardware interfaces 10
3.1.3.3 Software interfaces 10
3.1.3.4 Communication interfaces 10
3.1.4 System features 10

3.1.4.1 Clean HTML Tags 10
3.1.4.2 Remove Punctuation 10
3.1.4.3 Remove Stop words 11
3.1.4.4 Tokenization 11
3.1.4.5 Sentiment Classification 12
3.1.4.6 Handle Explicit Negation 12
3.1.4.7 Handle Explicit Contrast 13
3.1.4.8 Handle Sentimental Inconsistency 13
3.1.5 Other Non-functional Requirements 14
3.1.5.1 Performance requirements 14
3.1.5.2 Safety requirement 14
3.1.5.3 Security requirement 14
3.1.5.4 Software quality assurance 14
3.1.5.5 Business rules 14
Appendix A: Glossary 14
3.2 System design using DFD 15
3.3 Database Model using ER diagram 18
3.3.1 Entity Relationship diagram 18
Chapter 4: Conclusion 21
Chapter 5: References 22
Acknowledgments 24

List of figures

Figure 1: Level 0 DFD 15
Figure 2: Level 1 DFD 15
Figure 3: Level 2 DFD 16
Figure 4: Level 3 DFD 17
Figure 5: ER Diagram 18

List of Tables

Table 1: Comparative analysis of performance of different systems 5

Abbreviation, notation and nomenclature

1) DSA: Dual Sentimental Analysis
2) SVM: Support Vector Machines
3) PSDEE: Polarity Shift Detection, Elimination and Ensemble
4) CVS: Contextual Valence Shifters
5) TC-CVS: Term-Counting and Context Valence Shifters
6) BOW: Bag of words model
7) POS: Part-of-speech

Chapter 1

Introduction

With the recent trend of online shopping, buying and selling products online, analysts found that about 2.5 quintillion bytes of data is generated every day; with such a huge amount data being generated everyday it has been attracting many developers and this has leads to development of new branches like data mining and big data [1]. This data being generated has been influencing many people in some or the other way so there is a need to analyse the data forming a sub-branch better known as sentimental analysis. Sentiments can be said as thoughts, views or ideas expressed by an individual towards the product. In the same way many users express their views for the product leading to generation of large amount of data. These reviews can be useful to the other people for future references and help them to guide if they are interested in buying the following service, it also provide all the factual information about the product i.e. all the positive and negative.

In particular, online opinions have turned into a kind of virtual currency for businesses looking to market their products, identify new opportunities, and manage their reputations. Users just have to see the ratings which are generated by analysing the reviews given by other

users to that product and have to take his/her decisions. Such ratings are easily understandable by any user. But they don't give clear idea of how the product is. They are helpful only in the scenario where if any product is excellent or very poor. The scenario where product is average, star ratings prove bit confusing for any user and they don't give a clear views of what the other users think of that product. This is the reason why a rating should be provided based on the comments or reviews given by the users. If the ratings are given on the basis of the review they will give a clear idea of what the user thinks about the product and what positive or negative he/she has found in it.

Extracting sentiments from these comments is the first major task to be performed for analysing. While extracting these sentiments, the problems faced are handling polarity shifts of those reviews. Polarity of a review is termed as the presence of the positive or negative words in the review, when such positive or negative words occur in review polarity shift occurs. The biggest challenge is detection of such terms and handling them appropriately according to amount of positivity and negativity. Three types of polarity shifter have been proposed in earlier work and they are explicit negation, explicit contrast and sentimental inconsistency [2]. Explicit negation occurs when negative words or negators occurs in a statement and contribute to the polarity shift, Explicit contrast are the contrasting words in which user express opposite ideas and Sentimental inconsistency is the one in which the user expresses many ideas which may be opposite to each other and may lead to inconsistency of statement. So handling all 3 types of polarity shifter is one of the challenges.

The rest of the work done in report is as follows, In chapter 2 i.e. Review of Literature , the work done or proposed in various papers is discussed; the various aspects in which data can be analysed and processed which is followed by Comparative analysis in which different models are compared for various aspects such as Precision , Recall and Accuracy . In Chapter 3, a Report on Present Investigation is done which includes System Requirement Specification (SRS), System design and Database model. In SRS external interface Requirement, System features and non-Functional features are provided to the best of knowledge.

Chapter 2

Review of literature

There have been various methods and systems proposed for handling polarity shifts. Each proposed system gave certain accuracy in providing results by established algorithms. At present, numerous researchers have used many methods such as SVM, PSDEE model, Naïve Bayes method, contextual valence shifters, and maximum entropy classifier to classify the sentiments as positive, negative or neutral.

Xia and Feng Xu proposed dual sentiment analysis method for in which they trained the classifier and then the system would predict the sentiment [3]. They first created the reviews which would be sentimentally opposite to the original reviews; such reviews were called as reversed reviews. Then for training the classifier, they gave the pairs of these reversed and original reviews as input to the classifier. Various experiments performed, demonstrated that DSA model was effective for polar classification. Phan Thi Tuoi and Vo Ngoc Phu mainly focused on classifying the sentiments using Term-counting and Contextual Valence Shifters method [4]. This proposed method used a combined dictionary for recognizing and classifying the sentiments in the document.

Hui Song and Xiaoqiang Liu in their word proposed a system to extract the features from the reviews based on patterns [5]. The patterns of Part of speech tags and features from training corpus were extracted and applied into a pattern matching algorithm which then extracted the titles and opinion words from the reviews. The system created these patterns by splitting the sentences of the reviews (in their work: Chinese reviews) into sub-sentences based on punctuation for getting word segmentation and POS tags. Then from the POS patterns extracted, it would choose the best effective pattern to create pattern set. Their system achieved accuracy of 80% for extracting features. Takeru Yokoi and Roliana Ibrahim proposed a method of extracting the emoticons for representing the sentiment intention [6]. This emoticon extraction method was based on eye characters and string symmetry of emoticons where the target emoticons were classified into three categories; eastern emoticon (face with horizontal line), western emoticon (face with vertical line) and Japanese emoticon (multiple lines on the face). Valentina and Diego proposed a Sentilo model for detecting holders and topics of opinion sentences [7]. But this model was based on an assumption that the events and situations in the reviews would be the primary entities of an opinion which then was used for mapping the holder, topics and sub-topics.

Indhuja K in her work demonstrated Fuzzy logic for sentimental analysis of product reviews, which was used to compute fuzzy score of given reviews [8]. The logic involved the representation of sentences in form of parse trees using Stanford parser. The output of this parse tree was then fed into the fuzzy opinion mining model for obtaining the score. Then, based on the score, the reviews were classified as- very positive, positive, neutral, negative, very negative.

Li-Feng and Xi Liu focused on Chinese neologisms in the social media data in their work where neologism discovery methods were used for detecting neologisms [9]. After detecting the neologisms, their system determined the sentiment orientation based on TF-IDF method. But in their method, they achieved a very poor accuracy of 33% as they labelled many unknown words as “neologism” by mistake.

Rui Xia proposed a PSDEE model for handling polarity shifts. This model showed a different approach of modelling the polarity than the term counting method, i.e., it proposed a three stage model of first detecting polarity shifts, second eliminating the polarity shifts in negation and finally a polarity shifts ensemble model [2]. The system first divided the sentences into sub-sentences based on different types of polarity shift elements (in their work: negation, contrast, sentiment inconsistency), then their system would eliminate the negator from the negation sentences and then reviews would be ensemble. Compared to previous work, our system applies

the method of pre-processing the reviews, and then using sentiment classifier, classify the reviews into positive and negative.

Table 1: Comparative analysis of performance of different systems

Sr. No	Reference. No	Title	Authors	Datasets	Approach	Performance (accuracy)	Precision		Recall	
1.	[8]	Fuzzy Logic Based Sentiment Analysis of Product Review Documents	Indhuja K.	SFU review corpus	Sentiment classification	85.58%	84.86%		86.92%	
2.	[4]	Sentiment Classification using Enhanced Contextual Valence Shifters	Vo Ngoc Phu Phan Thi Tuoi	Internet movie dataset	Combination of Term counting and enhanced Contextual valence shifter method	84%	P 77.44%	N 61.10%	P 66.54%	N 72.95%
3.	[7]	Frame based detection of opinion holders and topics: A model and a tool	Aldo Gangemi Valentina Presutti	MPQA corpus Europarl corpus	Heuristic graph mining approach	78%	72%		64%	
4.	[9]	Associating Sentimental orientation of Chinese Neologism in social media data	Xi Liu Vincent Ng	3 million micro blogs posted by 2695 micro blog users	Metcalf's approach	33%	50.95%		92%	
5.	[10]	An approach based on Tree Kernel for opinion mining of online product review	Peng Jiang, Chunxia Zhang	Customer review data	Support Vector Machine, KNN, Perceptions model	89.56%	91.53%		93.1%	
6.	[11]	A syntactic approach for aspect based opinion mining	Chinsha TC, Shibily Joseph	Restaurant review from trip advisor	Approach to aspect level opinion mining using SentiWordNet, dependency parsing	78.04%	83%		89.25%	
7.	[5]	Extracting aspects and mining opinions in product review using supervised learning algorithm	A Jeyapriya, C.S. Kanimozhi Selvi	Customer review dataset (amazon,cnet,opinions)	Dynamic adaptive Support Apriori uses Naïve Bayes, Maximum entropy class, SVM	86.365%	Aspect extraction: 75% Sentiment orientation: 90%		85.71% 94.74%	

Chapter 3

Report on Present Investigation

3.1 Software Requirement Specification

3.1.1 Introduction

3.1.1.1 Purpose

The purpose of this document is to provide a detailed description of detecting and handling polarity shifts, extracting the sentiments and analysing the result.

3.1.1.2 Document Conventions

DFD = Data Flow Diagram

ER = Entity relationship

NLTK = Natural Language Tool-kit

3.1.1.3 Intended Audience

Primary readers of this document are the web researches, web designers and developers. This document is intended for the following:

Developers: In order to be sure they are developing the right project that fulfills requirements provided in this document.

Testers: In order to have an exact list of the features and functions that has to respond according to requirements and provided diagrams.

Documentation writers: To know what features and in what way they have to explain. What technologies are required, how the system will response in each user's action etc.

3.1.1.4 Project scope

The scope of the project is to provide a user-friendly product that extracts people's sentiments towards movies and its various features. In this project phase which aims to develop a filed prototype, we emphasize on handling polarity shifts in the movie reviews.

The project aims to:

1. Provide an accurate sentiment analysis result.
2. Smooth, efficient, reliable easy-to-use tool.
3. Detection of sarcasm.

3.1.1.5 References

The SRS template being used is taken from Karl Wiegers, author of Software Requirements.

3.1.2 Overall description

3.1.2.1 Product Perspective

Ratings for movies are generally provided on a scale of 10 or shown graphically like using stars. Such ratings don't give a clear understanding of why that movie is rated good or bad. The product is intended on providing a sentiment based classification into positive or negative using a pre-processed data.

3.1.2.2 Product functions

The product mainly focuses on handling the polarity shifts and providing ratings based on the sentiment classification on the reviews. Python environment will be used for pre-processing the data and then classify the sentiments for handling polarity shifts.

This system will provide the following functions:

1. Clean HTML Tags

Removing HTML tags is a part of pre-processing which shall clean the review.

2. Remove Punctuation

Removing Punctuation is an important part of data cleaning because it shall remove unwanted symbols and characters.

3. Remove Stop words

In this process the formally recurring words such as and, or, in, the, etc. shall be removed.

4. Tokenization

Tokenization is process in which tokens shall be extracted from a sentence, which adds sentiments to the sentence.

5. Sentiment Classification

This is the most important step of our tool. The training data shall be classified based on the sentiments as positive and negative and groups of limited number of reviews appending them will be formed, thus covering overall train data. Then, a 10 fold cross-validation will be performed on this data. The random forest classifier shall analyse the training data and accordingly predict the sentiment score for testing data.

6. Handling Explicit Negation

The most common type of polarity shift is explicit negator. In explicit negation, the negator shifts the polarity of the statement using some pre-defined negator.

7. Handling Explicit Contrast

In explicit contrast, some contrast indicators shift the polarity of previous phrase. Similarly, like explicit negator, some predefined contrast indicators shall be used.

8. Handling Sentimental Inconsistency

In Sentimental inconsistency, reviewer expresses the different perspectives in the same sentence.

3.1.2.3 User classes and characteristics

The web application is intended to be used by users having some basic knowledge of operating the system. The application would provide help to its users.

• Advanced end users: advanced users are those who have valuable input and feedbacks. Users who are more familiar with informative sites and can use our features efficiently. These valuable feeds will lead to enhancement of users' satisfaction.

3.1.2.4 Operating environment

The system interface will be implemented using Microsoft ASP.net technology. Also the system would require various SQL databases to be hosted which would hold large amount of data such as reviews provided by the opinion seeker as well as opinion provider. The system must be completely compatible with any browser that fully supports Microsoft ASP.NET technology.

3.1.2.5 Design and implementation constraints

The design and implementation constraints relating to the project are as follows:

(1) This system is domain specific; only applicable for movie domain.
(2) It is limited to providing output for reviews having textual format.
(3) It supports only English language.
(4) The system functions are implemented using free wares like Python and J-python.

3.1.2.6 User documentation

No user documentation is required for this system.

3.1.2.7 Assumptions and dependencies

It is assumed that the application will be developed using the ASP.NET technology.

It is assumed that the system will interface with a SQL Server 2000 database.

3.1.3 External Interface Requirements

3.1.3.1 User interfaces

User interface includes various forms and windows. The main window will consist of the registration form and login tools and help.The interface will visualize the features and functionalities listed in this document for this prototype:

• Field for searching movies.
• Push button for displaying reviews and ratings.
• Field for writing a review.
• Visualized representation for showing results.
• Help button.

3.1.3.2 Hardware interfaces

The solution makes extensive use of hardware device including:

• Windows user's computers.

3.1.3.3 Software interfaces

Other than the hardware interfaces specified, the software requirements are to support windows operating systems. For data gathering, IMDB is the only source we used.

3.1.3.4 Communication interfaces

Internet connection and a web browser are required in order to make use of several functions and to be executed such as searching, viewing and writing.

3.1.4 System features

3.1.4.1 Clean HTML Tags

3.1.4.1.1 Description and priority

All the HTML tags should be removed that are present in the review provided by the user; the cleaning of html tags is one of the methods in pre-processing.

Priority: High

3.1.4.1.2 Stimulus/Response Sequences

Stimulus: This module is activated after the user provides the review for the pre-processing method.

Response: All the HTML tags are removed from the review provided by the user.

3.1.4.1.3 Functional Requirements

REQ 1: HTML tags from the review must be removed.

3.1.4.2 Remove Punctuation

3.1.4.2.1 Description and priority

Punctuation like apostrophe, brackets, colon, comma, dash and many more shall be removed from the review provide by the user. Removing of punctuation and special characters also comes under the pre-processing method.

Priority: Medium

3.1.4.2.2 Stimulus/Response Sequences

Stimulus: This module is activated after the user provides the review for the pre-processing method.

Response: All the punctuation and numbers are removed from the review and the user gets cleaned review.

3.1.4.2.3 Functional Requirements

REQ 1: Punctuation like apostrophe, colon, brackets and many more must be removed from the review

3.1.4.3 Remove Stop words

3.1.4.3.1 Description and priority

Stop words such as a, the, on, and many more should be removed from the reviews using the NLTK package thus simplifying the reviews.

Priority: High

3.1.4.3.2 Stimulus/Response Sequences

Stimulus: This module is activated after the user provides the review for the pre-processing method.

Response: All the stop words and unnecessary words are removed from review that has been provided by user.

3.1.4.3.3 Functional Requirements

REQ 1: Stop words like most commonly used words must be removed

3.1.4.4 Tokenization

3.1.4.4.1 Description and priority

Reviews must be converted to lower case and split into individual words called as tokens which shall be the input for further processing.

Priority: High

3.1.4.4.2 Stimulus/Response Sequences

Stimulus: This module is activated after the user provides the review of movie for the pre-processing method.

Response: Text from review will be broken and converted into tokens.

3.1.4.4.3 Functional Requirements

REQ 1: After Tokenization we must get labels for untagged reviews with the help of tokens obtained from tagged reviews.

3.1.4.5 Sentiment Classification

3.1.4.8.1 Description and priority

In this phase, sentiments of the terms should be identified. This sentiments will be aggregated at sentence and then to document level. Analysis will be performed at document level.

Priority: High

3.1.4.8.2 Stimulus/Response Sequences

Stimulus: This module is activated after the user provides a review of the movie.

Response: The sentiments are identified and the sentiment polarity of each item is retrieved together with the percentage of positive and negative sentiment of the whole result.

3.1.4.8.3 Functional Requirements

REQ 1: The sentiments must be identified and classified into positive and negative sentiments from the whole.

3.1.4.6 Handle Explicit Negation

3.1.4.5.1 Description and priority

There are always some negator present in a sentence which add some excessive negativity to statement which are needed to be handled properly which shall be done using this technique.

Priority: High

3.1.4.5.2 Stimulus/Response Sequences

Stimulus: This module will be activated after Tokenization

Response: With the help of explicit negation, the negator present in the sentence can be detected and can be handled further.

3.1.4.5.3 Functional Requirements

REQ 1: The negator must be detected from the sentence of the review and the negator must be handled

3.1.4.7 Handle Explicit Contrast

3.1.4.6.1 Description and priority

They are always some contrasting words present in a statement which leads to addition of opposite ideas so at times it becomes difficult to rate a review.

Priority: High

3.1.4.6.2 Stimulus/Response Sequences

Stimulus: As this module is a part of handling polarity shift it is activated after Tokenization.

Response: This will help us in detecting contrasting terms.

3.1.4.6.3 Functional Requirements

REQ 1: The different contrast indicators like but, either, neither, nor and many more must be detected and handled

3.1.4.8 Handle Sentimental Inconsistency

3.1.4.7.1 Description and priority

When many opposite ideas are expressed in a statement, this leads to sentimental inconsistency.

Priority: High

3.1.4.7.2 Stimulus/Response Sequences

Stimulus: As this module is a part of handling polarity shift it is activated after Tokenization.

Response: This will help us detecting sentimental inconsistency of opposite views provided in a statement.

3.1.4.7.3 Functional Requirements

REQ 1: The different and opposite views in the same sentence must be detected and handled

3.1.5 Other Non-functional Requirements

3.1.5.1 Performance requirements

As for this prototype version, detection of system would be done for system crash, hang or an operating system error occurred. Also, detecting the performance of the system in terms of the efficiency of integration of the different components.

3.1.5.2 Safety requirement

For the safety requirements, nothing but an operation of weekly backups for the data base should take place.

3.1.5.3 Security requirement

There are no specific security requirements, anyone can access the web application but only authorized users can submit a review and view the ratings.

3.1.5.4 Software quality assurance

• Reliability:

The solution should provide reliability to the user that the product will run with all the features mentioned in this document are available and executing perfectly. It should be tested and debugged completely. All exceptions should be well handled.

• Accuracy:

The solution should be able to reach the desired level of accuracy.

3.1.5.5 Business rules

The business rule of the project is that, reviews would be taken from the users but their personal identity details would not be revealed.

Appendix A: Glossary

1) Negation: contradiction or denial of something
2) Tokens: Tokens can be words, symbols, or other meaningful elements obtained by breaking up a stream of texts. This process of breaking up is known as tokenization.
3) Stop-words: These are the natural language words used very frequently like “a”, “an”, “as”, and similar words.

3.2 System design using DFD

A data flow diagram is one of the most common methods used for graphically representing a data through an information system. A DFD shows what information is provided as input to and output from the system.

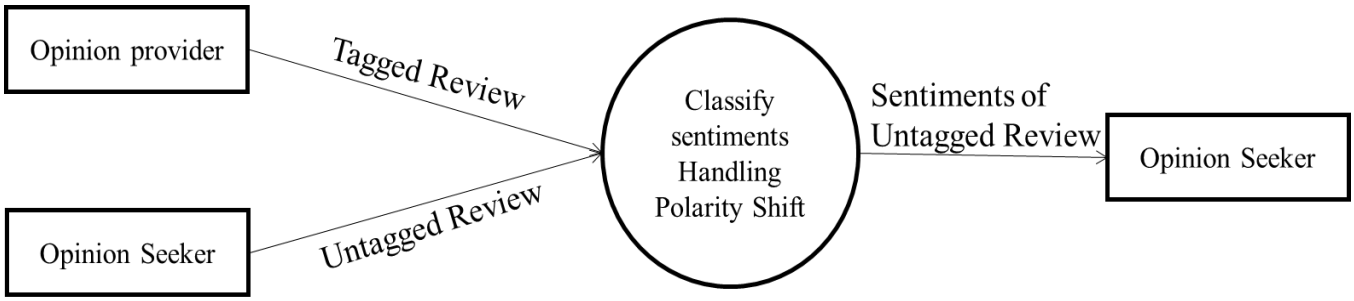

Figure 1: Level 0 DFD

Figure 1 level 0 also known as Context level DFD, gives the overview of whole system in which the external entities like the opinion provider and seeker provide tagged and untagged reviews, which are processed and sentiments of the untagged reviews are provided to the opinion seeker.

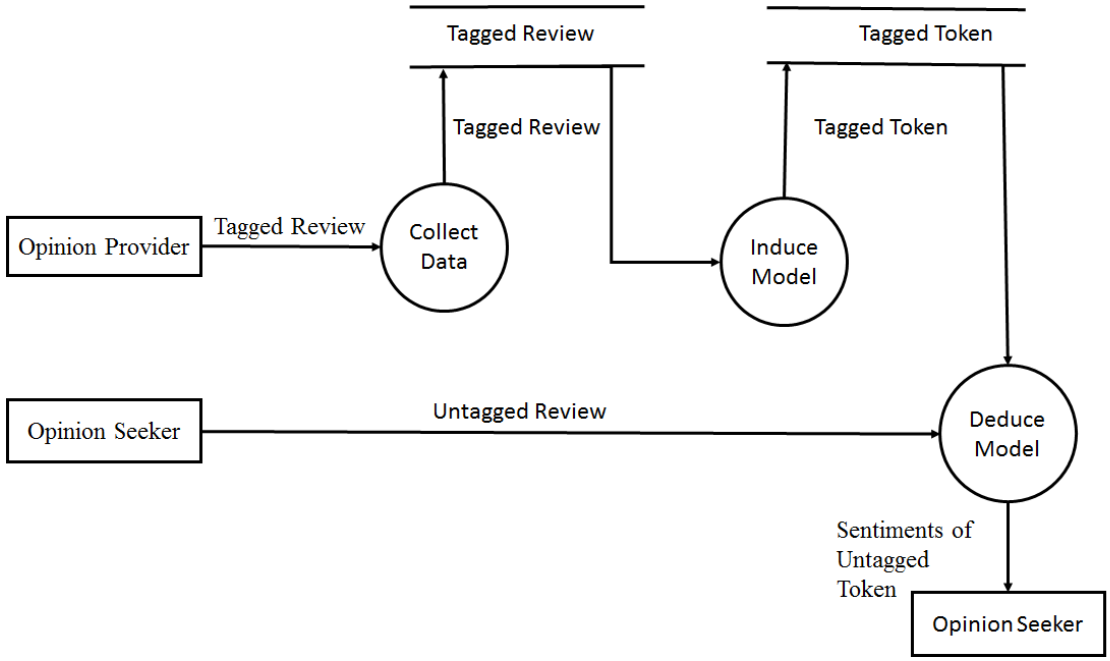

Figure 2: Level 1 DFD

Figure 2 represents collection of data, induce and deduce model processes and databases of tagged reviews and tagged tokens. It is elaboration of level 0 DFD.

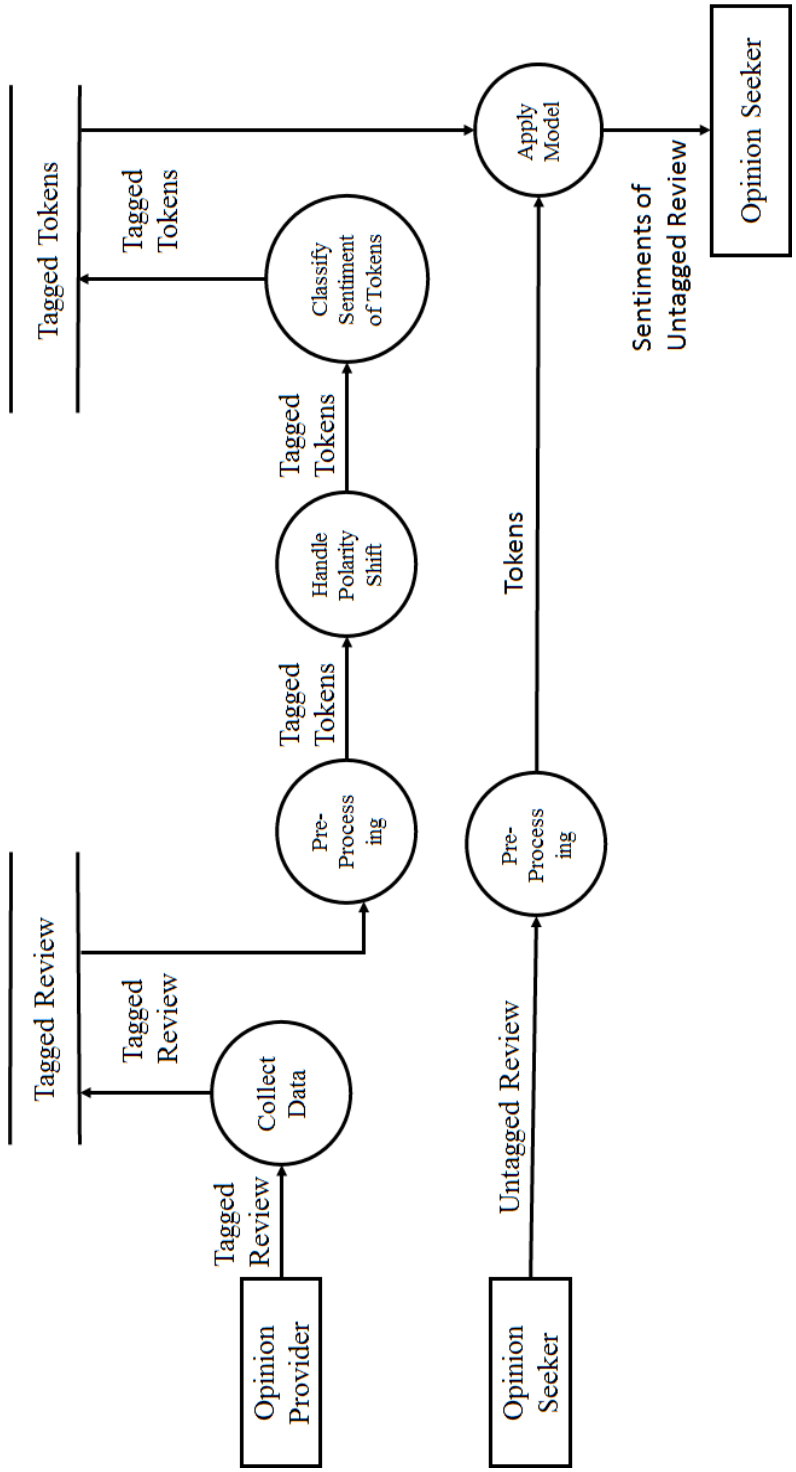

Figure 3: Level 2 DFD

In figure 3, the induced model is further elaborated into three processes- “pre-processing”, “Handle polarity shift” and “classify sentiments of token”. Whereas, the deduced model is elaborated into “pre-processing” and the model being implemented.

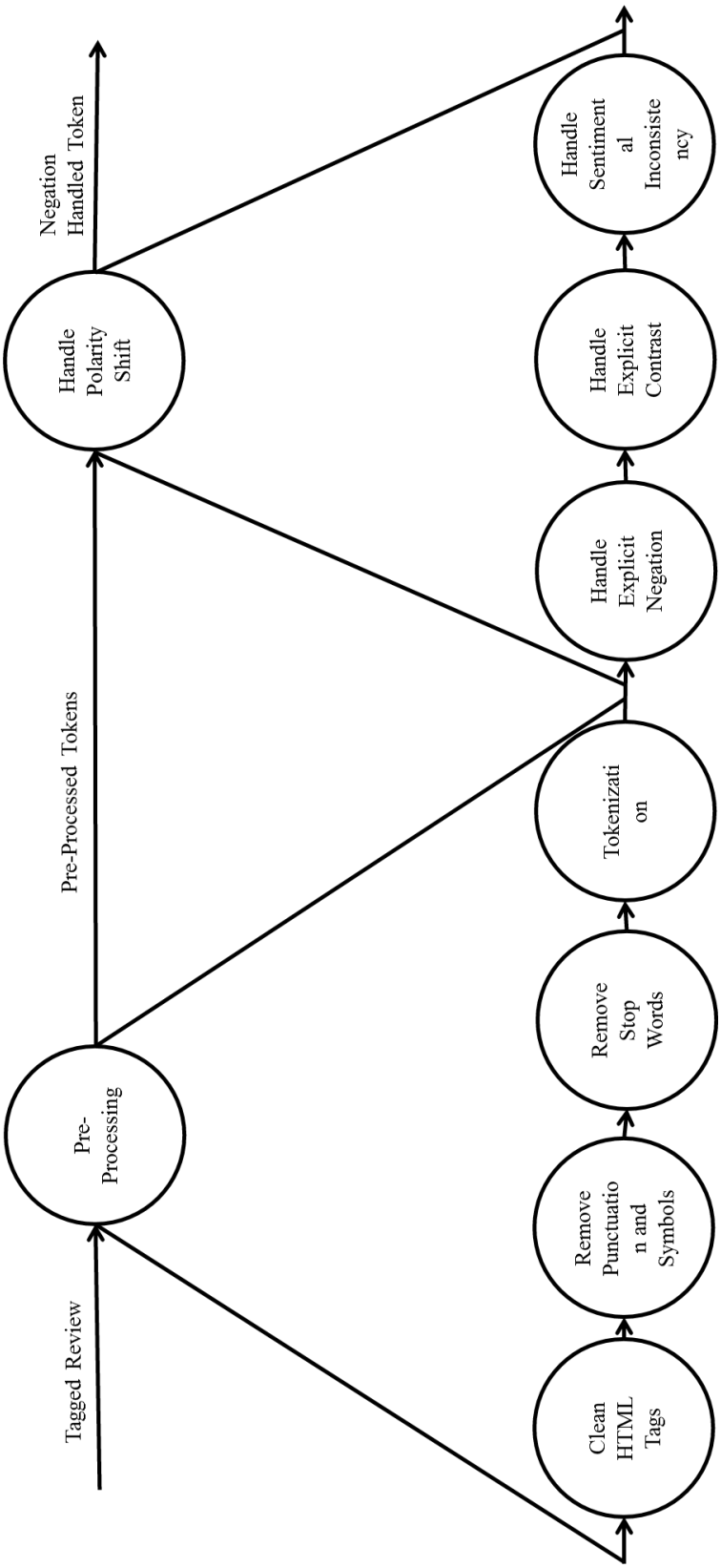

Figure 4: Level 3 DFD

Figures 4 represent elaboration of pre-processing and handle polarity shift processes. Pre-processing is further elaborated into cleaning HTML tags, removing punctuation, symbols and stop words, tokenization. Handle polarity shift is further elaborated into handling explicit negation, handling explicit contrast and handling sentiment inconsistency.

3.3 Database Model using ER diagram

3.3.1 Entity Relationship diagram

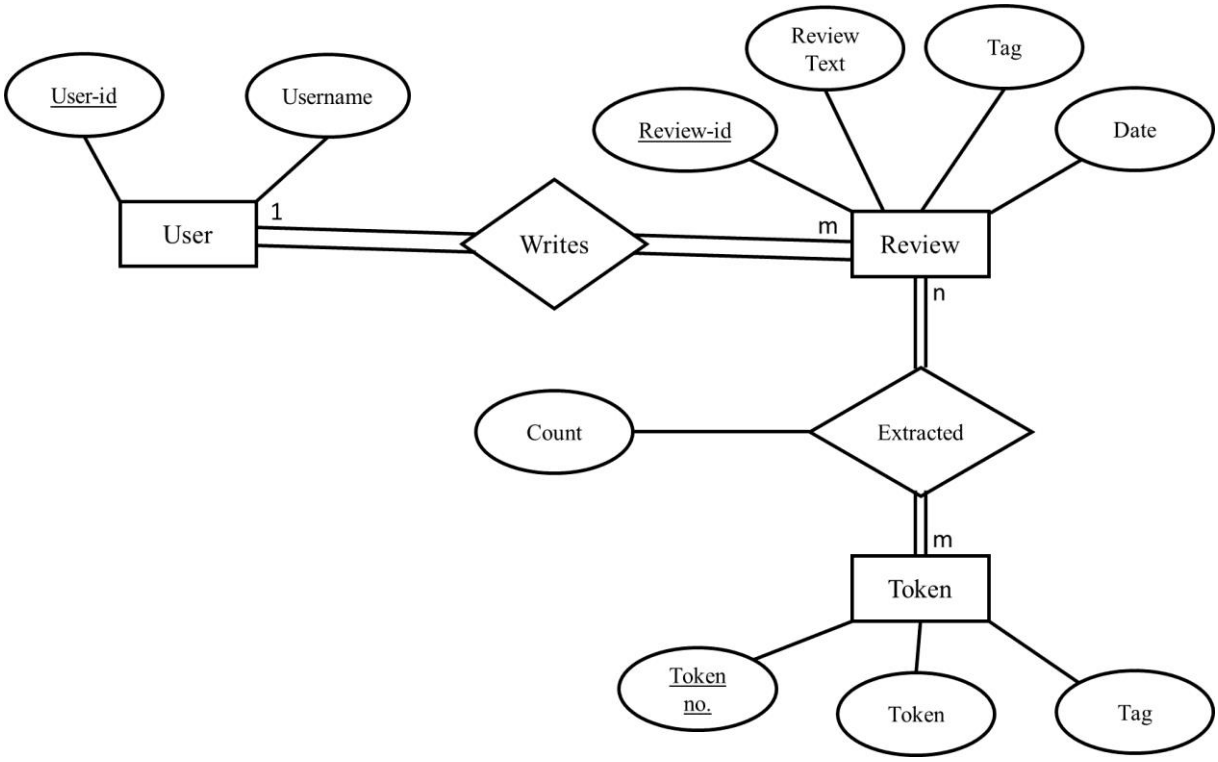

Figure 5: ER Diagram

In Entity Relationship Diagram, various attributes, entities and their relationship is discussed. **Figure 5** demonstrating the entity relationship diagram for project, consists of three entities; user, reviews and tokens. Each of these entities has certain attributes. User writes a review from which tokens are extracted and counted.

Step 1: Map Regular Entities

User (User-id, Username)

Review (Review-id, text, date, tag)

Tokens (Token no., token, tag)

Step 2: Map weak Entities

There are no weak entities. So the schema remains same.

Step 3: Map relationship with cardinality m: n

User (User-id, Username)

Review (Review-id, text, date, tag)

Tokens (Token no., token, tag)

Extracted (Review-id, Token no., count)

Step 4: Map relationships with cardinality m: 1 or 1: m

User (User-id, Username)

Review (Review-id, text, date, tag, User-id)

Token (Token no., token, tag)

Extracted (Review-id, token no.)

Step 5: Map relationship with cardinality 1: 1

There are no such relationships. So the schema remains same.

Step 6: Map N-any relationship

There are no such relationships. So the schema remains same.

Step 7: Map multivalued attribute

There are no multivalued attribute. So the schema remains same.

Step 8: Map External ER features

There are no External ER features. So the schema remains same.

Final Schema

User (User-id, Username)

Review (Review-id, text, date, tag, User-id)

Token (Token no., token, tag)

Extracted (Review-id, token no.)

Schema definition

User

Attribute	Data type	Constraints
Uid	Integer(10)	Primary key
Uname	Char(20)	-

Review

Attribute	Data type	Constraints
Review_id	Integer(10)	Primary key
Text	Char(500)	-
Date	Integer(10)	-
Tag	Char(10)	-
User _id	Integer(10)	Foreign key reference from user

Token

Attribute	Data type	Constraints
Token_no	Integer(100)	Primary key
Token	Char(10)	Primary key
Tag	Char(10)	-

Extracted

Attribute	Data type	Constraints
Review_id	Integer(10)	Foreign key references from Review Primary key
Token_no	Integer(10)	Foreign key references from Token Primary key

Chapter 4

Conclusion

Polarity shift is one of the hindrances which affect the output accuracy of many rating systems available online. For handling the polarity shift, a model is proposed consisting of Pre-processing, Tokenization and negation handling. Initially evaluation was performed for data cleaning on mentioned data-set such as removing HTML tags , removing stop words and special symbols, and with the help of Random forest classifier we had done 10 fold cross validation which gave us an average accuracy of 83.417%. While rating for any movie, our system will mainly focus on handling polarity shift extending the scope of negation modifier handling. Apart from modification and polarity shift handling, other task for sentiment classification such as pre-processing and tokenization will be performed in traditional way. Thus, the system eventually aims to provide a better accuracy as compared to previous works.

Chapter 5

References

[1.] “Every Day Big Data Statistics – 2.5 Quintillion Bytes of Data Created Daily,” *Every Day Big Data Statistics – 2.5 Quintillion Bytes of Data Created Daily*. [Online]. Available: http://www.vcloudnews.com/every-day-big-data-statistics-2-5-quintillion-bytes-of-data-created-daily/. [Accessed: 14-Sep-2016].

[2.] R. Xia, F. Xu, J. Yu, Y. Qi, and E. Cambria, “Polarity shift detection, elimination and ensemble: A three-stage model for document-level sentiment analysis,” *Information Processing & Management*, vol. 52, no. 1, pp. 36–45, 2016.

[3] R. Xia, F. Xu, C. Zong, Q. Li, Y. Qi, and T. Li, “Dual Sentiment Analysis: Considering Two Sides of One Review,” *IEEE Trans. Knowl. Data Eng. IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 8, pp. 2120–2133, Jan. 2015.

[4] V. N. Phu and P. T. Tuoi, “Sentiment classification using Enhanced Contextual Valence Shifters,” *2014 International Conference on Asian Language Processing (IALP)*, 2014.

[5] Hui Song and Xiaoqiang Liu, “Extracting product features from online reviews for Sentimental analysis”, College of Computer science and Technology, Doghua University, China, pp. 745-750, 2011.

[6] T. Yokoi, M. Kobayashi, and R. Ibrahim, “Emoticon Extraction Method Based on Eye Characters and Symmetric String,” *2015 IEEE International Conference on Systems, Man, and Cybernetics*, 2015.

[7] A. Gangemi, V. Presutti, and D. R. Recupero, “Frame-Based Detection of Opinion Holders and Topics: A Model and a Tool,” *IEEE Comput. Intell. Mag. IEEE Computational Intelligence Magazine*, vol. 9, no. 1, pp. 20–30, 2014.

[8] K. Indhuja and R. P. C. Reghu, “Fuzzy logic based sentiment analysis of product review documents,” *2014 First International Conference on Computational Systems and Communications (ICCSC)*, 2014.

[9] L.-F. Huang, X. Liu, and V. Ng, “Associating sentimental orientation of Chinese neologism in social media data,” *2015 IEEE 19th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 2015.

[10] P. Jiang, C. Zhang, H. Fu, Z. Niu, and Q. Yang, “An Approach Based on Tree Kernels for Opinion Mining of Online Product Reviews,” *2010 IEEE International Conference on Data Mining*, 2010.

[11] C. T. C and S. Joseph, “A syntactic approach for aspect based opinion mining,” *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*, 2015.

Acknowledgments

We are sincerely thankful to our guide Ms. Kranti Ghag for giving us an opportunity to work under her guidance and for providing excellent support.

We convey our special acknowledgements and gratitude to our review committee.

We are also grateful to Ms. Swati Deshpande, the HOD of Information Technology department for guiding and supporting.

We would also like to thank our honourable principal Dr. Bhavesh Patel, for motivating and supporting us.

Tanmay Shukla	B.E. - 5 - 63	----------
Vidhan Thakur	B.E. - 5 - 73	----------
Kandaswamy Thevar	B.E. - 5 - 74	----------
Deepak Vishwakarma	B.E. - 5 - 77	----------

Date: