Opinion Mining
of
Online Customer Reviews

Patlammagari Gowtamreddy

Department of Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela – 769 008, India

Opinion Mining
of
Online Customer Reviews

Dissertation submitted in

May 2014

to the department of

Computer Science and Engineering

of

National Institute of Technology Rourkela

in partial fulfillment of the requirements

for the degree of

Master of Technology

by

Patlammagari Gowtamreddy

(Roll 212CS1362)

under the supervision of

Korra Sathya Babu

Department of Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela – 769 008, India

dedicated to my Parents and inspiring guide...

Computer Science and Engineering

National Institute of Technology Rourkela

Rourkela-769 008, India. www.nitrkl.ac.in

Dr. Korra Sathya Babu

Professor

May , 2014

Certificate

This is to certify that the work in the thesis entitled **Opining Mining of Online Customer Reviews** by **Patlammagari Gowtamreddy**, bearing roll number 212CS1362, is a record of an original research work carried out by him under my supervision and guidance in partial fulfillment of the requirements for the award of the degree of *Master of Technology* in **Computer Science and Engineering**. Neither this thesis nor any part of it has been submitted for any degree or academic award elsewhere.

Korra Sathya Babu

Acknowledgement

This dissertation, though an individual work, has benefited in various ways from several people. Whilst it would be simple to name them all, it would not be easy to thank them enough.

The enthusiastic guidance and support of *Prof.Korra Sathya Babu* inspired me to stretch beyond my limits. His profound insight has guided my thinking to improve the final product. My solemnest gratefulness to him.

I am also grateful to *Dr.S.K.Rath* for his ceaseless support throughout my research work. My sincere thanks to *Prof.S.K.Jena* for his continuous encouragement and invaluable advice.

Many thanks to my comrades and fellow research colleagues. It gives me a sense of happiness to be with you all. Special thanks to *Jitendra Kumar Rout and Pramit Mazumdar* whose support gave a new breath to my research.

Finally, I am grateful to all my friends for continuous motivation and encouragement. Last but not the least to my family having faith in me and always supporting me.

Patlammagari Gowtamreddy

Declaration

I Patlammagari Gowtamreddy of Computer Science and Engineering with Roll no 212CS1362 hereby declare that the project submitted by me is solely of my work and is not copied from any other source where ever may available and It has not been previously submitted for any academic degree. I had verified my thesis report through Turnitin software for plagiarism. All sources of quoted information have been acknowledged by means of appropriate references.

If in future my work was found to be plagiarized from any other persons work, then in that situation I alone will be responsible for it.

Date: June 2014

Patlammagari Gowtamreddy

NIT Rourkela

Abstract

Customer Opinions play a very crucial role in daily life. When we have to take a decision, opinions of other individuals are also considered. Now-a-days many of web users post their opinions for many products through blogs, review sites and social networking sites. Business organizations and corporate organizations are always eager to find consumer or individual views regarding their products, support and service. In e-commerce, online shopping and online tourism, its very crucial to analyze the good amount of social data present on the Web automatically therefore, its very important to create methods that automatically classify them. Opinion Mining sometimes called as Sentiment Classification is defined as mining and analyzing of reviews, views, emotions and opinions automatically from text, big data and speech by means of various methods. In this thesis we are going to see how Apriori frequent item set mining algorithm can be used for mining reviews from online reviews those are posted by customers. Our main theme is to create a system for analyzing opinions which implies judgement of different consumer products.

Keywords: Opinion Mining, Sentiment Classification, Apriori Algorithm, SentiWordNet, Min-max Normalization, Frequent Words, Online Reviews.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

In this chapter, we give a brief introduction about what is Opinion Mining and how it can be performed. Opinion Mining or Sentiment classification involves building a system to make use of reviews posted by the users and opinions that are expressed in blogs and forums as comments and reviews and sometimes may be as tweets about the product.

1.2 Opinion Mining

Opinion Mining is the science which combines techniques of computational linguistics and information retrieval and is concerned with the opinions expressed rather than topics in the text. Opinions are written on many things example a product, a topic, an individual, etc. In opinion mining task we identify the orientation of opinion by the holder towards any object which may be a collection of features or components or attributes.

Opinion mining [7, 8] might be valuable in a few ways. For instance, in advertising, it tracks and judges the achievement rate of a commercial crusade or launch of new item, focus prevalence of items and administrations with its forms additionally let us know

about demographics which like or hate specific characteristics. Case in point, a survey may be around a computerized Polaroid may be comprehensively positive, yet be particularly negative about how overwhelming it is. The seller gets overall picture of general opinion than studies and centre gatherings, if this sort of data is indentified in a methodical manner.

1.3 Sentiment Analysis or Sentiment Classification [9]

What do other individuals think has dependably been a vital component in choice making methodology. Sentiment Analysis or Sentiment Classification is the methodology to naturally focus the sentiments communicated in a bit of plain content utilizing some regular automated preparing systems. To be specific, term Sentiment is exceptionally wide and it constitutes feelings, opinions, dispositions, particular encounters, and so forth. In this theory, we speak just about the opinions communicated in writings which are composed in texts which are written in human readable natural language, in social media.

Sentiment analysis is a procedure for following the views of the clients around a specific item or subject. Sentiment analysis, which is likewise called opinion mining, includes in building a framework to gather and look at opinions about the item made in blog entries, remarks, audits or tweets. Sentiment analysis might be helpful in a few ways. Case in point, in showcasing it helps in judging the achievement of a notice crusade or new item dispatch, figure out which forms of an item or administration are prominent and even recognize which demographics like or aversion specific attributes.

1.4 Customer Produced Content [1, 2]

The web has all of a sudden changed the way how individuals express their perspectives and opinions. They can now post audits on different sites, take an interest in discourse on different discussions, compose a site depicting their experience, redesign their status on social sites like Linkedin, Facebook, Google+ and etc. This information on audit sites, exchange gatherings, websites, and interpersonal organizations might be combined and called as customer produced substance. Each of these customers created substance has their special property. In this thesis, we focus on product reviews given by the customers in various blogs, review sites and forums.

1.4.1 Review sites

There are millions of sites where customers can compose views, reviews and opinions about products, movies, and so on. A number of online shopping sites like Amazon, Flipkart allow their clients to compose reviews on the product. These reviews impact the choice of the new client who needs to purchase the item. Sites like Epinions, is an accumulation of reviews and opinions from different products like vehicles, books, hardware, software and so forth. There are a numerous sites like Rotten Tomatoes, Imdb which simply have film reviews of wide sort and dialect space. These sites are extremely helpful from the point of view of opinion mining.

1.4.2 Blogs

A Blog is a page where an individual or gathering of customers record opinions, data, and so on all the time. Web blogs are composed on numerous various subjects like governmental issues, games, travel and even items. Be that as it may, the nature of the content produced from these sources is by generally noisy and poor. There are numerous web entrances like

blogspot, wordpress, and so on where any customer can make a record and post a web blog. Numerous item and news sites likewise have a committed websites segment where limited approved customers or users can post online reviews.

1.4.3 Dataset

We can take Dataset from many websites like Amazon, myntra and other online shopping sites for performing Opinion Mining process which includes analyzing the reviews that are posted by the customers or users on various products.

1.5 Motivation

Interest for sentiment analysis and opinion mining is expanding numerous fold. Customers use web as a verbal exchange to express their choices focused around the opinions communicated by others. The social media associate the whole world and are one of the explanations behind data over-burden on the web. Twitter has hundreds of millions users who handle very nearly half a million tweets for every day, normal of thousands of tweets for every second. These tweets are posted on different dialects not simply English. These bewildering volumes of upgrades call for a mechanized analysis of this text. There are numerous structures in which customer produced substance is distributed on Internet. This inspired us to devise systems to handle each of these various types of information and concentrate some helpful data.

1.6 Objectives

Given an object and a collection of reviews on it, our objectives are

- Extract Nouns, Adjectives, Verbs and Adverbs on the remaining reviews by using dictionary approach.

- Identify frequent words by using Apriori frequent item set mining Algorithm.
- Perform Sentiment Analysis on the frequent words using SentiWordNet.
- Provide visualization.

1.7 Thesis Organisation

The rest of the thesis contains various chapters as mentioned below:

Chapter 2, we discuss about the terminology and different methods used in opinion mining. Then we outline several previous research efforts on opinion mining that attempts to develop a method of extracting opinions.

Chapter 3, we present our approach towards opinion mining from customer reviews.

Chapter 4, we give the results we got for the data set containing customer reviews which are posted on online review sites and we made the comparison between the results we got for Apriori frequent item set mining algorithm using SentiWordNet and for star ratings that we got from customer reviews.

Chapter 5, finally we conclude this thesis with future work.

Chapter 2

Introduction to Opinion Mining

A large portion of the current content handling systems work with authentic data. The web holds enormous volumes of opinionated content. Web users express particular emotions and opinions on very nearly anything at review sites, blogs, and forums and so on. This important data is freely accessible for internet clients. The substantial gathering of opinions on the Web makes it extremely tough to get helpful data effectively. Perusing all reviews and emotions to settle on an educated choice is a much time taking task. Perusing distinctive and potentially even conflicting opinions composed by diverse commentators may make organizations, users and customers more confused. These needs have made another line of examination on mining user and customer opinions, which is called opinion mining.

Opinion mining is the part of study that dissects individual opinions, sentiments, assessments, mentality, and feelings from written text. It has pulled in a number of analysts from distinctive areas of exploration including NLP, information mining, machine learning, phonetics, and even social science. In the current chapter, we first examine need for opinion mining and afterward characterize the terminologies utilized within this study. In the coming

sections, we talk about general opinion mining tasks and present a concise survey of the current and related deals with each one.

2.1 Need for Opinion Mining

With the growth of online social networking sites, for example, forums, review sites, blogs, and micro blogs, the enthusiasm towards opinion mining has expanded essentially. Today online opinions have transformed into a sort of virtual profit for business organizations looking to market their items, recognize new trends and deal with their position. Many organizations are currently utilizing opinion mining systems to track customer inputs in online shopping sites and review sites.

Opinion mining [10] is additionally helpful for organizations to analyze customer opinions on their products and features. While product attributes are clearly mentioned, discovering the primary cause behind low profit needs much focus on all the more on individual customer views on such characteristics. Opinion mining is an amazing method for taking care of numerous business trends identified with deals administration, status management, and advertising. Additionally, organizations may have the capacity to perform pattern prediction in deal by following customer perspectives.

2.2 Opinion Mining Terminology [3, 4]

In this segment we define the various terms used in the opinion mining.

Fact: A fact is that which has truly happened or is really the case.

Opinion: An opinion is a view or judgement formed about something, not necessarily based on fact or knowledge.

Subjective Sentence: A sentence or a text is subjective or opinionated if it actually indicate ones feelings.

Objective Sentence: An objective sentence indicates some facts and known information about the world.

Item: an individual article or unit, especially one that is part of a list, collection, or set.

Review: A review is a text containing a sequence of words thats has opinions of customer for a specific item. A review may be subjective or objective or both.

Known Aspects: Known aspects are default aspects provided by the certain website for which users separately give ratings.

Sentiment: Sentiment is a polarity term that implies to the direction in which a concept or opinion is expressed. We use sentiment in a more specific sense as an opinion about an aspect. For example, excellent is a sentiment for the attribute 'battery life' in the sentence "This mobile has excellent battery life".

Opinion Phrase: An opinion phrase is a pair of head term and modifier. Usually the head term is a candidate aspect, and the modifier is a sentiment that expresses some opinion

towards this aspect.

Opinion Polarity: Opinion Polarity or Subjectivity Orientation denotes the polarity expressed by the user or customer in terms of numerical values.

- **Polarity:** Polarity is a two way orientation scale. In this, a sentiment can be either positive or negative.
- **Rating:** Most of the reviewing websites use star ratings for expressing polarity, presented by stars in the range from 1 to 5 which are called ratings.

Overall Rating: All the online shopping websites ask customers to give an overall rating for the product that they already bought mentioning the overall quality of the used item.

Part-of-Speech (POS) Tag: POS tagging is very useful in Opinion Mining process. When we need to analyse a document or a sentence first we have to extract the subjective information from the document or that particular sentence. POS tagging helps us in getting subjective words like Nouns, Verbs, Adverbs and Adjectives. After extracting these words we can perform various actions on these and we can come to a conclusion.

2.3 Steps in Opinion mining approach

In this segment we show a survey of the current and related tasks in brief at opinion mining proposed in the existing methods. For a thorough survey, we sort opinion mining procedure into three general classes, yet before we talk about our arrangement, we introduce current order constructions experienced in the writing. Pang et al. [11] group the significant issues of opinion mining into three classes: sentiment polarity identification, subjectivity

detection, and joint topic-sentiment analysis. Some experts additionally characterizes three mining methods for opinionated content in his book. He further grows this classification in his handbook as: sentiment and subjectivity characterization, aspect based opinion mining, sentiment analysis of similar sentences, opinion search and data retrieval, and opinion spam identification. At last, in his latest book he characterizes three general classifications for opinion mining tasks: document level, sentence level, and phrase level.

There is one more interesting study in the opinion mining tasks classification called feature level. We discuss about feature level opinion mining in brief in this section.

- Document level opinion mining
- Sentence level opinion mining
- Phrase Level opinion mining
- Feature Level opinion mining

2.3.1 Document-Level Opinion Mining [5]

Document level tasks primarily concerns with classification issues where the available document has to be arranged into a set of predefined classes. In subjectivity analysis a document is divided as subjective or objective. In sentiment analysis, a document can be classified as positive or negative or neutral depending upon the polarity of subjective information that

is present in the document. Opinion quality and support evaluation makes decision whether an opinion is useful or not and opinion spam identification groups and divides opinions as spam and not spam.

Subjectivity Analysis

Subjectivity Analysis [12, 13] refers to finding whether the given document makes an opinion or not. To be precise, whether a document or text is objective or subjective. We consider this problem generally as a classification problem. Many of the current methods uses supervised learning, even though we have few unsupervised methods. One of the works in this area given by Wiebe et al. [14] does subjectivity analysis using the naive Bayesian classifier. Succeeding research uses other learning algorithms for finding subjective text. Future research has been mainly focused on developing automated process for subjectivity analysis. One of the tough tasks in subjectivity classification is the human effort involved in labelling training examples as subjective or objective.

Sentiment Analysis

Sentiment classification in default takes the given review is subjective and expects to discover the general sentiment of user in the content. Suppose, we have a product review, it figures out if the review is of positive polarity or negative polarity. Sentiment classification, as other way to subjectivity examination, does not typically require manual exertion for training the data. Preparing information utilized as a part of sentiment classification are generally online product reviews that already been marked by reviewers. Current works preferably focuses

on supervised learning approach to sentiment classification. We have three existing machine learning methods namely naive Bayes, Maximum Entropy classification, and Support Vector Machine (SVM) to classify movie reviews as positive polar or negative polar. We infer that the standard machine learning methods perform far better than human produced approaches. We have few unsupervised methods for dividing reviews.

Opinion Quality and support evaluation [15]

The issue of consequently assessing the support of online reviews has additionally pulled in expanding consideration. Most past works endeavour to foresee the supportiveness of reviews by utilizing a set of selected features, like literary and social features and taking in a capacity of these features for foreseeing survey accommodation. Literary features incorporate features that are focused around content facts, for example, length of the survey, the normal length of a sentence, rate of things or descriptive words, and so forth. Social features, then again, are identified with the making of the survey and are concentrated from his social setting, for example, the amount of past reviews by the creator, in-degree and out-level of the creator in the informal community, past normal rating for the creator, and so forth.

Opinion Spam identification

Web spam is very well known to most of the individuals. In the view of opinion, we have a comparative spam issue. The development of customer created substance inspires more individuals to discover and read opinions on the Web. Positive opinions can bring about profit and money related addition or acclaim for the maker or merchants, and negative opinions,

then again, can have the opposite effect on them. This effect additionally gives great motivating forces for composing spam reviews called as opinion spam. Spam opinions attempt to deliberately misdirect customers by giving undeserving positive or negative opinions to some target things so as to push the thing as well as by giving undeserving negative opinions to some different things to harm their position.

Spam identification could be planned as a characterization issue with two classes: spam and no spam. The principle assignment is to discover the situated of viable information features for building the model. Jindal et al. [16, 17] distinguish some text based and social features for taking in a relapse model to gauge the likelihood of each one review being a spam. In distinctive systems for distinguishing spam reviews utilizing conduct of reviewers are proposed. There are likewise a few deals with recognizing opinion leader, and finding gathering spammers for reviews.

2.3.2 Sentence-Level Opinion Mining [6]

The problem at this measure everything refers to sentences. In opinion data extraction and recovery and opinion question answering, sentences are generally placed and positioned focused around some criteria. Opinion rundown intends to select a set of sentences which outline the opinion all the more exactly. At long last, opinion mining in relative sentences incorporates recognizing similar sentences and concentrating data from them.

Opinion Search and Retrieval

Conventional Web quest is exceptionally essential for online customers. Opinion retrieval [18] will be likewise of incredible utilization. Looking the customer produced substance on

the Web empowers review readers to discover opinions on any topics. Opinion data extraction methods are for the most part issued to discover popular opinion on a specific thing or a part of the thing. For instance, to discover general opinion on a multimedia mobile, a customer may raise the problem "mobile battery life".

With respect to the classification task, web search tools generally rank WebPages focused around correctness and certain score. The presumption is that the top positioned pages hold sufficient data to fulfil the customers data need. In any case, this suspicion is not genuine in the concept of opinions. The top positioned documents just speak to the opinions of few persons not people in general. Current positioning techniques use distinctive criteria to reflect general customer opinion. The system proposed in the behavioural model of shoppers utilizing budgetary methodology for positioning items. Likewise in different tasks, review quality, content detail, customer inputs are granted as metrics of positioning.

Opinion visualization [19]

The objective of opinion visualization is selecting entire sentences to make a short passage outline. Provided for a few reviews, a review outline system yields content comprising requested sentences. Existing methodologies concentrate on selecting sentences so that the outline incorporates vital data of the reviews.

Opinion Question Answering

Question Answering is one of the useful step in opinion mining. Opinion question answering

process [20] gives a way to answer the decision based queries using the customers opinions about the specific product. There is lot of difference between traditional question answering and opinion question answering process. Opinion question answering is rather complicated when compared to traditional question answering process. For instance consider the question which is the best android mobile available in the market today? and some questions do target on semantic opinion like You suggest me to buy this camera or not? Opinion question answering is generally much time taking process and can be considered as virtual. Business Organizations compare and classify the polarity of reviews and opinion question and answers. The methods that are already in process creates a method that classifies between subjective and objective at document level and then we get polarity of those by using various sentiment classification approaches.

Opinion mining in similar sentences

Opinion mining in similar sentences in similar sentences refers to comparing two products of similar category or sometimes we may compare only certain attributes of the different products. Opinion mining in comparative sentences [21] demands the use of words like best, better, larger, smaller and so on. These approaches are considered as classification problems and generally require some machine learning techniques and training a classifier.

2.3.3 Phrase-Level Opinion Mining

Phrase level mining [5] came into picture because document level mining and sentence level mining approaches cant find accurately what actually users likes and they does not like. Phrase level opinion mining looks for sentiments on features of products.

Aspect-Based Opinion Mining

In the recent days, lot of researchers are showing a great interest in aspect based opinion mining [22]. We have a few approaches to retrieve aspects from comment and reviews. Many of these methods use full text comments and reviews, where as a few approaches have taken advantage of comments and reviews those are in sequential pattern.

2.3.4 Feature Level Opinion Mining

Feature level opinion mining comes into picture when a customer or user looking for feedback of certain feature or attribute of a product rather than total feedback of the product. We see many customers interested in only certain features of some products rather than the whole product like some people look for a mobile that has excellent battery life and they are not concerned with other features like camera clarity, music clarity and so on. In situations like mentioned in this section feature level opinion mining helps a lot for extracting polarity information for a particular feature or attribute from a product.

Extract object features

In this step, we extract product features from comments and reviews. We may follow any approach for doing this, but frequently used method is to extract nouns and noun phrases there by getting actual features of the product.

Determine polarity of opinions on features

After extracting product features from reviews and comments using nouns and noun phrases, we perform semantic analysis on the resulting features which gives information about the each and every feature which customer liked and which he did not like.

Group feature synonyms

In some scenarios, we may end up a situation like we have various synonyms for same attribute and feature from a product. By grouping those feature synonyms, our task becomes easier for performing opinion mining process. Suppose if a customer has written a review mentioning that "this mobile can stand by for 2 days without battery recharging", this statement actually refers to hidden feature called 'battery life.

Chapter 3

Proposed Approach

In this section, we discuss about proposed approach for performing opinion mining and what techniques and Algorithms we are going to use to perform Opinion Mining and Sentiment Analysis for getting useful information from online customer reviews. Our methodology includes the following steps:

- Extract Nouns, Adjectives, Verbs and Adverbs from the customer reviews by using dictionary approach [23].

- Identify frequent words by using Apriori frequent item set mining algorithm [24, 25, 4].

- Perform Sentiment Analysis on the frequent words using SentiWordNet [26, 27, 28].

- Use Min-max normalization [29, 30] for normalizing the polarity of values that we got using SentiWordNet and for normalizing the values of star ratings that we have taken

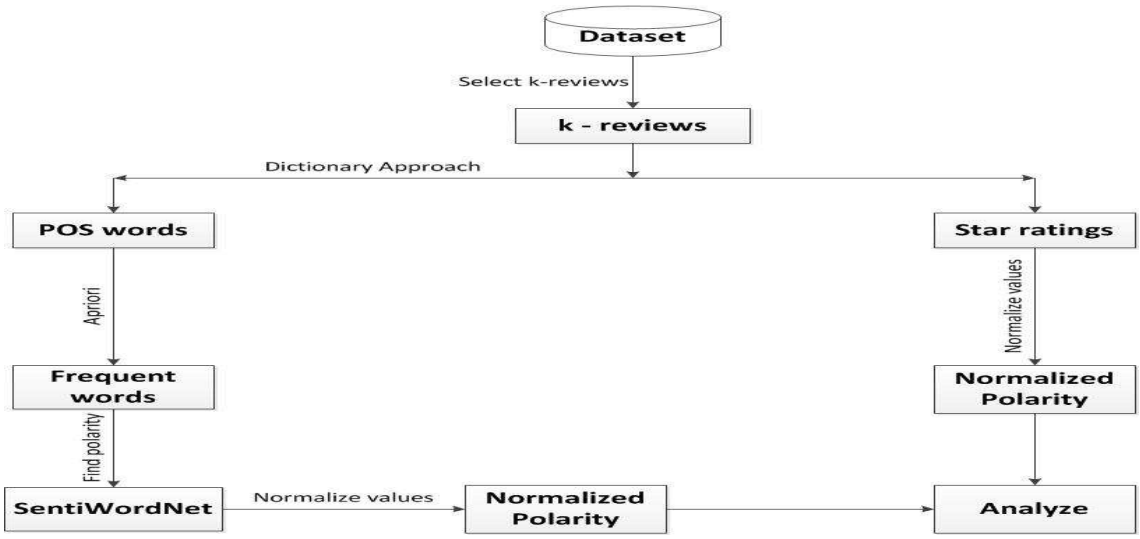


Figure 3.1: Proposed Approach

from the customer reviews.

- Provide visualization.

The following figure shows an overview of proposed approach 3.1.

3.1 Extract Nouns, Adjectives, Verbs and Adverbs

To perform this step, we are using dictionary approach. After this step we get all the Nouns, Adjectives, Verbs and Adverbs that are present in the customer reviews file. We use these words for getting the frequent words in the next step.

3.2 Identify the frequent words by using Apriori Algorithm.

In this step, we get frequent words by using the various Apriori frequent itemset mining algorithm.

3.2.1 Apriori Algorithm

The following are the steps involved in Apriori Algorithm.

Assume for C_k and L_k

C_k denotes candidate itemset of k size

L_k denotes frequent itemset of k size

Important steps of algorithm are:

1) Initially get frequent set L_{k-1}

2) Join step: get C_k by doing cartesian product of L_{k-1} with itself

3) Those itemsets which are of size (k-1) and those are not frequent should not be a subset of a frequent itemset of size k, so those should be removed

4) Finally frequent set L_k has been achieved

3.3 Sentiment Analysis on the frequent words using SentiWordNet

In this step we perform Sentiment Analysis on the frequent words that we got from Apriori Algorithm by using SentiWordNet. It provides a value for each and every word.

Sentiment Analysis deals with the usage of automated techniques for anticipating the introduction of subjective substance on text reviews or comments, with usage in various fields that includes recommendation system and advertising, user intelligence and opinion retrieval. Sentiwordnet is an opinion vocabulary and can be considered as extended from the Wordnet database where each one term is connected with numerical scores demonstrating positive and negative sentiment data. This examination shows the consequences of applying the Sentiwordnet lexical asset to the issue of automated sentiment arrangement of customer film reviews or comments.

3.4 Min-max Normalization

Min-Max normalization is the technique of taking data calculated in its own units and converting it to a value between 0 and 1

We use normalization because star ratings values lies between 1 to 5 and word polarity of SentiWordNet values lies between -1 and +1

Suppose we have some n rows with five variables, A, B, C, D and E, in the data. We use variable B as an example for understanding the normalization concept in the calculations below. All the other variables in the rows are normalized in the similar way.

The normalized value of B_i for variable B in the ith row is calculated as:

$$normalised(B_i) = \frac{B_i - B_{min}}{B_{max} - B_{min}} \tag{3.1}$$

where

B_{min} = the least value for variable E

B_{max} = the highest value for variable E

If $B_{max} = B_{min}$ then Normalized (B_i) is set to 0.5.

Chapter 4

Implementation and Results

This section contains results of our proposed Opinion Mining approach.

4.1 Dataset

The following picture shows the set of customer reviews for Bellagio hotel that is situated in Las Vegas, USA. In the given file each row denotes one review with first column being customer id, second column refers to star rating given by the customer for the product and third column denotes the review written by the customer for the product.

The following 4.2 shows the set of Nouns, Adjectives, Verbs and Adverbs in the text file. We search for these words in the data set and extract them from the text file containing text file.

4.2 Extracting Nouns, Adjectives, Verbs and Adverbs

In the following figure 4.3 we see the extracted words using POS tagging from text file containing customer reviews. The following figure represents a bar chart between number of reviews and number of words extracted for different number of reviews.

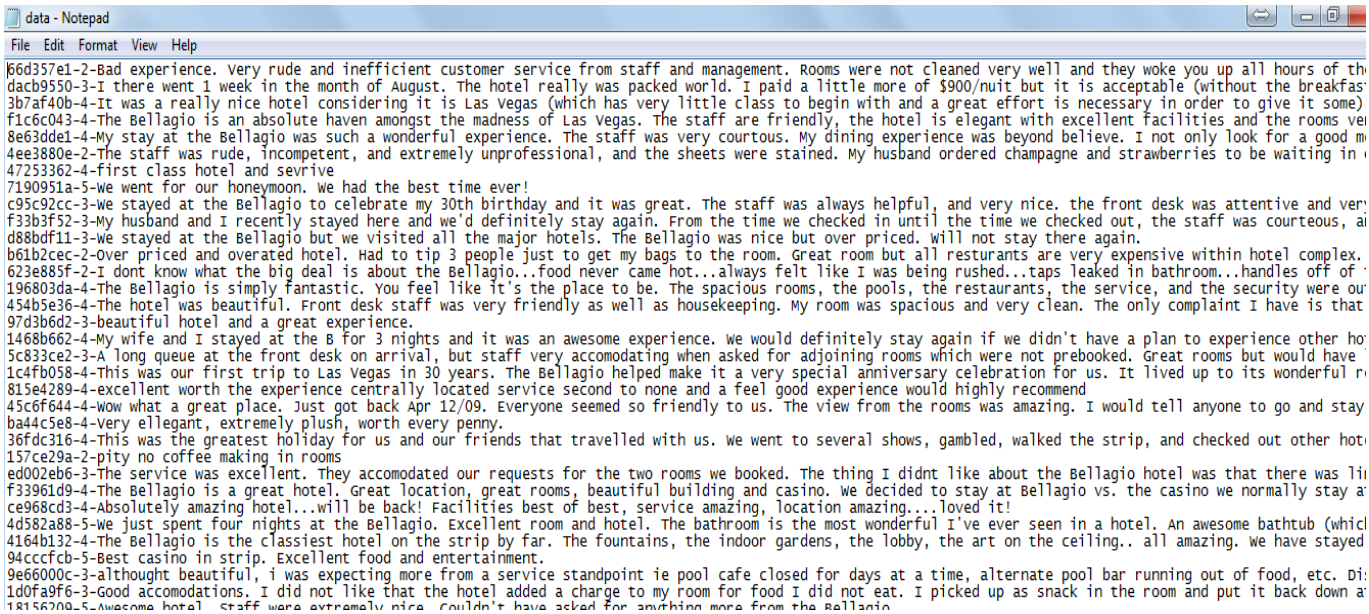


Figure 4.1: dataset1

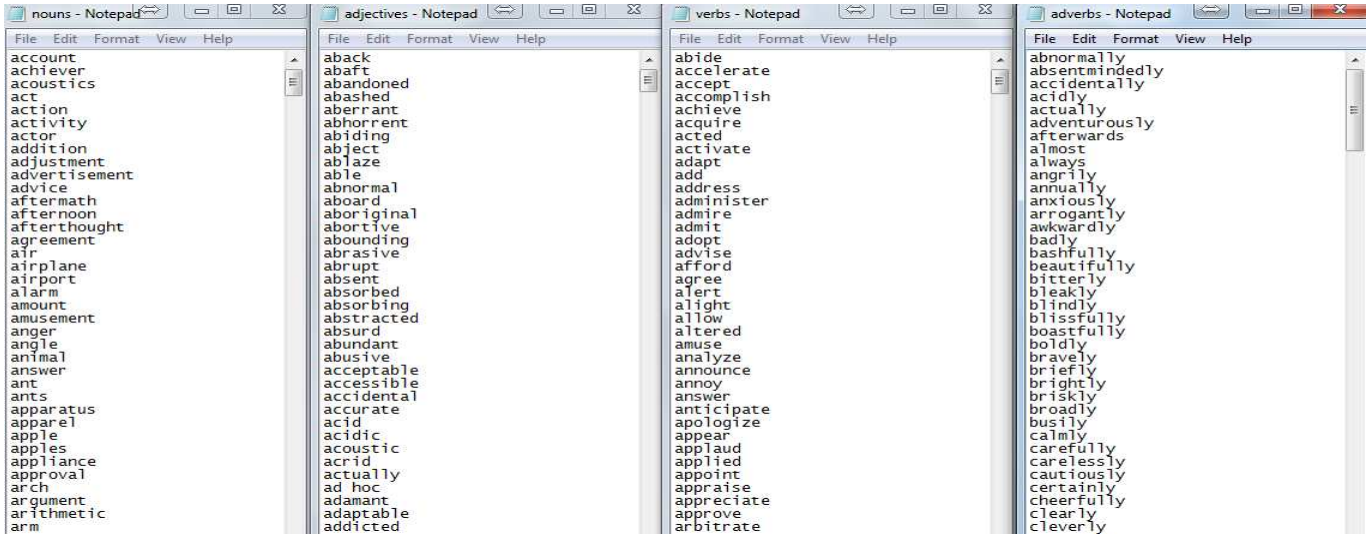


Figure 4.2: dataset2

4.3 Extracting frequent words using Apriori Algorithm

This figure 4.5 shows the frequent words we got using Apriori Algorithm from the set of words.

```
Python 2.7.5 Shell
File Edit Shell Debug Options Windows Help
Python 2.7.5 (default, May 15 2013, 22:43:36) [MSC v.1500 32 bit (Intel)] on win32
Type "copyright", "credits" or "license()" for more information.
>>> ============================== RESTART ==============================
>>>
hotel inefficient service staff bellagio nice wonderful beautiful experience air art back bath bed believe birthday book breakfast building can card change class compa
ny condition daughter day desk dinner door drink experience fact floor food front furniture glass hand holiday hot hour idea level look love maid marble middle money m
onth morning night order part person place price regret room shop smell smoke star street surprise system thing time trip trouble turn use view waste way week wish ama
zing awesome best better big closed complex even excellent far first four free hard hot last like long necessary next nice past polite right rude same second slow smal
l special strong superb two waiting whole young especially extremely not truly be begin come cost deal do drink eat eliminate feel fit get give go hold know leave let
love maintain make mistake order own pay perfect put queue recommend return say schedule see service sit slide smell strip tell thank think trip understand use write
>>>
```

Figure 4.3: Extracting Nouns, Adjectives, Verbs and Adverbs

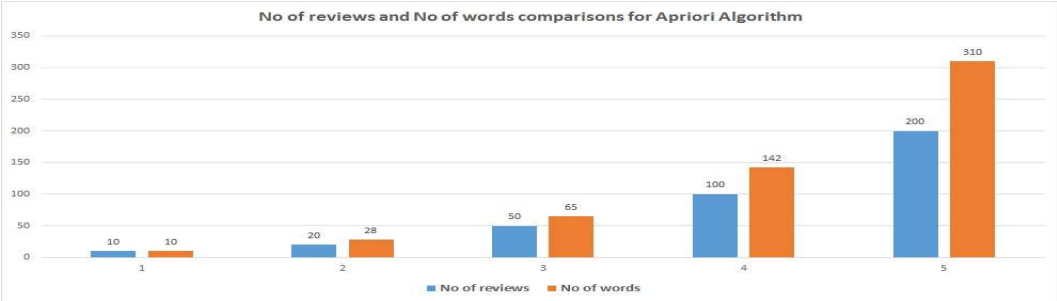


Figure 4.4: Extracting Nouns data

4.4 Sentiment Analysis on the frequent words using SentiWordNet

The following figure 4.6 show how we can use SentiWordNet for finding the polarity of a word.

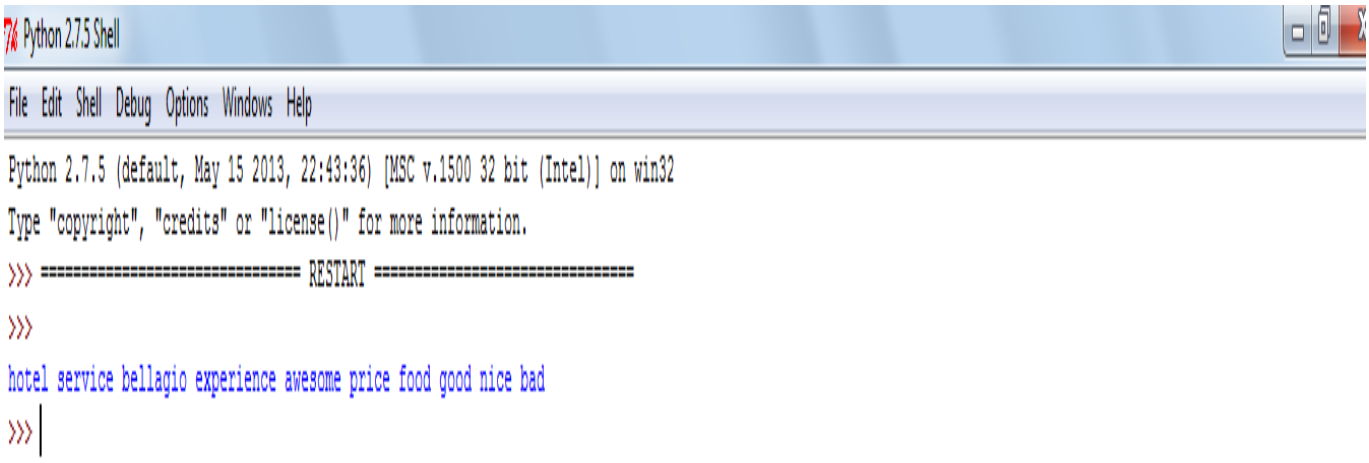
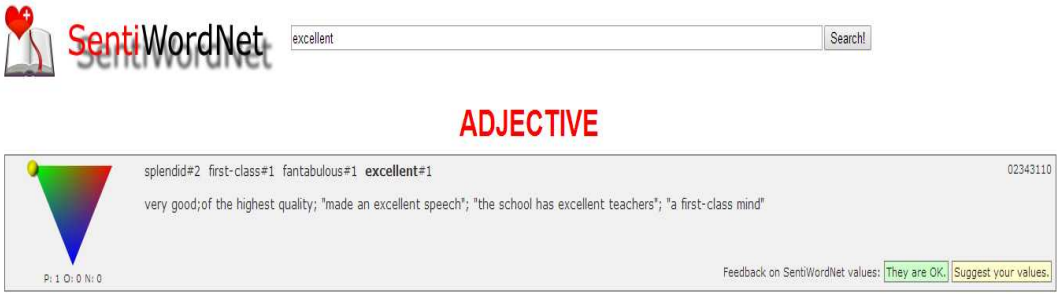


Figure 4.5: Extracting frequent words using Apriori Algorithm



(a) Sentiment analysis-1

(b) Sentiment analysis-2

Figure 4.6: Using SentiWordNet for finding polarity of a word

4.5 Min-max Normalization

We have star ratings value ranges between 1 to 5 and word polarity of SentiWordNet values ranges between -1 and +1. So for mapping both the values on a scale of 0 to 1, we use Min-max normalization.

The following table shows the normalized polarity values for star ratings taken from customer reviews and for SentiWordNet for different number of reviews 4.1.

No.of Reviews	**Star rating**	**Sentiwordnet**
10	0.6	0.48
20	0.57	0.5
50	0.52	0.47
100	0.42	0.39
200	0.38	0.36

Table 4.1: Normalized Polarity values of Star ratings and Sentiwordnet scores

4.6 Comparison between star rating and Opinion Mining process

The following graph shows how the normalized polarity values gets varied for star ratings values that are taken from customer reviews and for Opinion Mining process that uses Apriori algorithm and SentiWordNet for different number of reviews.

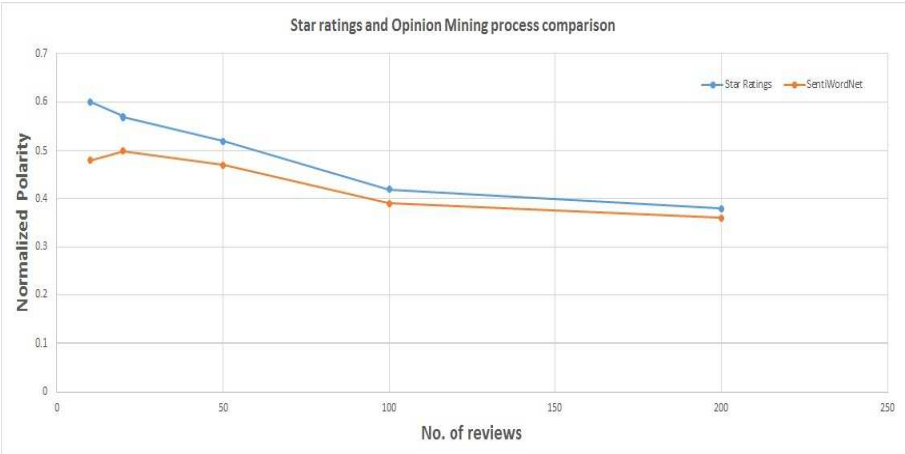


Figure 4.7: Comparison between star rating and Opinion Mining process

Chapter 5

Conclusion and Future Work

Opinion mining has become a fascinating research area due to the availability of a huge volume of user-generated content in review sites, forums and blogs. Opinion mining has applications in a variety of fields ranging from market research to decision making to advertising. With the help of opinion mining, companies can estimate the extent of product acceptance and can devise strategies to improve their product. Individuals can also use opinion mining tools to make decisions on their buying by comparing competitive products not just based on specifications but also based on user experience and public opinions. In this thesis we have shown how Apriori frequent item set mining algorithm can be used in Opinion Mining process. We have seen how normalized polarity varies for the values taken from customer reviews star ratings and from values of opinion mining process that uses Apriori algorithm and SentiWordNet.

Here, we see a few but important challenges for text analytics tasks like opinion mining. It is very difficult to distinguish between objective and subjective information, generally opinion words also occur in objective sentences, so it is very tough to handle these challenges. Many times we see customers posting the reviews in the blogs or forums

with a lot of spelling mistakes which our dictionary cannot catch them and resulting in less accuracy of desired output. Most times of the times we see many spam blogs and spam reviews posted by the users. If we consider these reviews for performing Opinion Mining, we may get deviated from our desired results. So, lot of work has to be done in this field for identifying spam blogs, considering spelling mistakes and for other challenges.

Bibliography

[1] Luís Sarmento, Paula Carvalho, Mário J Silva, and Eugénio de Oliveira. Automatic creation of a reference corpus for political opinion mining in user-generated content. In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 29–36. ACM, 2009.

[2] John Krumm, Nigel Davies, and Chandra Narayanaswami. User-generated content. *IEEE Pervasive Computing*, 7(4):10–11, 2008.

[3] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

[4] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

[5] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics, 2005.

[6] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*, 2010.

[7] Andrea Esuli. Automatic generation of lexical resources for opinion mining: models, algorithms and applications. In *ACM SIGIR Forum*, volume 42, pages 105–106. ACM, 2008.

[8] Richa Sharma, Shweta Nigam, and Rekha Jain. Supervised opinion mining techniques: A survey.

[9] Akshat Bakliwal. Fine-grained opinion mining from different genre of social media content. 2013.

[10] Samaneh Moghaddam and Martin Ester. Aspect-based opinion mining from online reviews. In *Tutorial at SIGIR Conference*, 2012.

[11] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics, 2004.

[12] Ellen Riloff, Janyce Wiebe, and William Phillips. Exploiting subjectivity classification to improve information extraction. In *Proceedings of the National Conference On Artificial Intelligence*, volume 20, page 1106. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005.

[13] Stephan Raaijmakers and Wessel Kraaij. A shallow approach to subjectivity classification. In *ICWSM*, 2008.

[14] Janyce M Wiebe, Rebecca F Bruce, and Thomas P O'Hara. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 246–253. Association for Computational Linguistics, 1999.

[15] Anindya Ghose and Panagiotis G Ipeirotis. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *Knowledge and Data Engineering, IEEE Transactions on*, 23(10):1498–1512, 2011.

[16] Nitin Jindal and Bing Liu. Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 219–230. ACM, 2008.

[17] Nitin Jindal and Bing Liu. Review spam detection. In *Proceedings of the 16th international conference on World Wide Web*, pages 1189–1190. ACM, 2007.

[18] Wei Zhang, Clement Yu, and Weiyi Meng. Opinion retrieval from blogs. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 831–840. ACM, 2007.

[19] Lun-Wei Ku, Yu-Ting Liang, and Hsin-Hsi Chen. Opinion extraction, summarization and tracking in news and blog corpora. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, volume 100107, 2006.

[20] Hoa Trang Dang and Karolina Owczarzak. Overview of the tac 2008 opinion question answering and summarization tasks. In *Proc. of the First Text Analysis Conference*, 2008.

[21] Murthy Ganapathibhotla and Bing Liu. Mining opinions in comparative sentences. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 241–248. Association for Computational Linguistics, 2008.

[22] Charu C Aggarwal and ChengXiang Zhai. *Mining text data*. Springer, 2012.

[23] Xiaowen Ding, Bing Liu, and Philip S Yu. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 231–240. ACM, 2008.

[24] Raffaele Perego, Salvatore Orlando, and P Palmerini. Enhancing the apriori algorithm for frequent set counting. In *Data Warehousing and Knowledge Discovery*, pages 71–82. Springer, 2001.

[25] Christian Borgelt and Rudolf Kruse. Induction of association rules: Apriori implementation. In *Compstat*, pages 395–400. Springer, 2002.

[26] Bruno Ohana and Brendan Tierney. Sentiment classification of reviews using sentiwordnet. In *9th. IT & T Conference*, page 13, 2009.

[27] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422, 2006.

[28] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204, 2010.

[29] Anil Jain, Karthik Nandakumar, and Arun Ross. Score normalization in multimodal biometric systems. *Pattern recognition*, 38(12):2270–2285, 2005.

[30] Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1367. Association for Computational Linguistics, 2004.