OPINION MINING SYSTEM OF CUSTOMER REVIEWS BY USING FEATURE EXTRACTION

NANDAR MOH MOH LWIN

M.C.Sc **JULY 2023**

OPINION MINING SYSTEM OF CUSTOMER REVIEWS BY USING FEATURE EXTRACTION

BY

NANDAR MOH MOH LWIN

B.C.Sc

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Computer Science

(M.C.Sc)

University of Computer Studies, Yangon

JULY 2023

ACKNOWLEDGEMENTS

Firstly, I would like to express my gratitude and sincere honor to all persons who gave me a lot of help and support for the successful completion of this thesis.

Secondly, I would like to express my thanks to **Prof. Dr. Mie Mie Khin**, Rector of the University of Computer Studies, Yangon, for her kind permission to do this thesis.

Thirdly, I am sincerely grateful to my supervisor, **Dr. Tin Tin Htar**, Associate Professor, Department of Information Technology Support and Maintenance of the University of Computer Studies, Yangon, for her encouragement, invaluable guidance and strong technical support from the initial level to the final level to deepen the understanding of the subject.

Fourthly, my heartfelt thanks and respect go to **Dr. Si Si Mar Win** and **Dr. Tin Zar Thaw**, Professors, Faculty of Computer Science, for their kind administrative support as a course coordinator throughout the development of the thesis.

Fifthly, I am indebted to **Daw Mya Thandar**, Associate Professor, Department of English of the University of Computer Studies, Yangon, for her guidance and editing the thesis from the language point of view.

Sixthly, I am especially grateful to my parents for their unending support, encouragement, love, and invaluable support to fulfill all my wish.

Seventhly, I would like to thank my teachers especially domain experts and friends from the University of Computer Studies, Yangon, for their thoughtful and help for my seminar presentation. I wish to acknowledge with the deep gratitude for the valuable guidance received from my teachers who gave advice for the study.

Lastly, I deeply pay respect and blessings to everyone who has supported me in every respect during the completion of my thesis.

ABSTRACT

Due to the dramatic improvement of ecommerce, web sources which are important for both potential customers and service providers rapidly emerge in prediction and decision purposes. Opinion mining techniques become popular to automatically process customer reviews by extracting features and user opinions expressed over them. To overcome the task of manual scanning through the large amount of one-by-one review, people have interested to automatically process the various reviews and to provide the information which is useful for customers and service providers. By applying dependency relations, it can properly identify the semantic relationships between features and opinions of each review. It can find the numeric score of all the features using SentiWordNet. This system is intended to collect customer reviews from tourism field and then extract the related features and opinions to rate the services. Finally, it can rank each agency according to the final result of each review sentence. In this thesis, Standard Parser is used to generate the features, opinions and the dependency relations for each trip review at the preprocessing. The two methods of features extraction such as frequency-based feature extraction and dependency grammar-based feature extraction are used to extract the most relevant trip features. Moreover, SentiWordNet 3.0 is used to get the positive score and negative score for each trip feature and then the system calculates the total weight of the trip review by using these numeric scores. The objective of the system is to rank the travel agencies according to the final weight of each travel agency that is collected by adding the total weight of the trip reviews for that agency. Therefore, the system implements efficiency and effectiveness in opinion mining to express the reviewer's opinion and feeling for next customers' trip plans by using features extraction. In this system, Tourism Reviews are applied as the case studies to identify what elements of an agency affect sales most and what are the features the customer like or dislike so that trip managers and agency owners can target on those areas. The system is developed using Java language and MySQL to build the database.

CONTENTS

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Nowadays, there are many customers who use the services via websites. Many e-commerce companies often ask their customers to review the purchased products and related services. Manufacturers can target product characteristics by reading reviews from online websites. Therefore, manufacturers can identify which features of a product have the greatest impact on sales and which features customers like or dislike. As e-commerce popularity increases, the number of customer reviews is growing rapidly. Opinion mining is a research area both in natural language processing and in the information search community. It aims to find subjective information that could be relevant to users in order to obtain useful information in many applications.

A significant number of websites, blogs and forums allow customers to post comments on a variety of products and services (tripadvisor.com, amazon.com, etc.). These reviews are important resources to help customers make purchasing decisions. The main task of opinion mining based on a collection of customer opinions is to extract customer opinions and predict the direction of sentiment. Opinion mining algorithms are used to track and manage customer opinions through the orientation of mining issues and customer opinions online. Certain keywords mentioned in customer reviews are extracted along with the keywords. The proposed system evaluates the services provided by each travel agent. The system aims to find opinions by using customer opinions which are available on the web.

Today, most travel agencies conduct opinion polls to analyze customer attitudes, opinions, and emotions. Currently, clients use social media to share positive and negative experiences about travel agencies. Sentiment analysis can identify positive reviews and negative reviews that show the strengths and weaknesses users who write online. Sentiment analysis is based on algorithms that use natural language processing to classify writing elements as positive, neutral, or negative. This algorithm is designed to distinguish positive and negative words such as "great", "amazing", "terrible", and "unhappy". Due to the complexity of the language,

sentiment analysis must face at least some problems. The main problem with sentiment analysis tools is the contrasting conjunction. That is, a sentence is composed of two conflicting words (with positive and negative meanings). Moreover, sentiment analysis has to face poor grammar and incorrect spelling.

1.1 Overview of the System

There are several websites which are available on the website such as www.tripadvisor.com will permit the customers to contribute their views about hotels, tourist place, shopping website etc. Travel websites like TripAdvisor and Travelocity are nowadays important tools for travelers when deciding which hotels to stay in, which restaurants are good and tourist attractions to visit and which travel agencies provide the excellent services and the worst services to customers. Travel agencies with a customer-based service are the areas where multiple factors may impact customer sentiment. The large amount of opinionated data may provide both customers and agency owners' valuable information.

Opinion mining is used to identify and extract subjective information form user reviews and then to determine the sentiment of the text. Feature extraction is a technique to identify and extract trip features of the travel agencies such as hotel, food, staff, service, location, accommodation, transportation, internet, guide and so on. It includes an agency search, where customers can find agencies based on the features from customer reviews. This system can analyze the massive amount of agency information written by customers and can help agency owners or managers to analyze their organization. In this system, reviews from TripAdvisor (https://www.tripadvisor.com/) are collected and analyzed by using SentiWordNet and their results are summarized by the five selected features. These features are the most frequently features that are mentioned by the customers. This system firstly collects the trip reviews and puts them in the review corpus by given inputs. The output is the ranking features by summarizing the reviews.

1.2 Objectives of the Thesis

The objectives of the Thesis are as follows:

- To extract and analyze the agency rating from customer reviews

- To identify the features, so as to get more features about each agency review and cover all features of that agency
- To be convenient for choosing a better travel agency becomes much easier
- To save time by evaluating the large set of customer reviews
- To help travel agencies for tracing the rate of their services and merchandise

1.3 Organization of the Thesis

This thesis consists of five chapters. **Chapter 1** represents introduction to opinion mining and objectives of the thesis and organization of the thesis. **Chapter 2** deals with the background theory about the data mining, opinion mining, sentiment analysis by expressing the three levels of this, POS Tagging, Stanford Parser, Dependency Relation for Feature- Opinion Mining, Extracting Features, Frequency-based Feature Extraction, Term Frequency- Inverse Document Frequency, Dependency Grammar-Based Feature Extraction, Extraction Opinion, SentiWordNet 3.0 and Ranking the Features. **Chapter 3** is the step by step calculation of the system. **Chapter 4** gives the implementation of the system with fully implemented system flowchart, class diagram, sequence diagrams and entity relationship diagram, table design of the system and implementation of the system. Finally, **Chapter 5** describes the conclusions, advantages and limitations of the system.

1.4 Related Work

Nowadays, there has been a wide range of research based on customer reviews from studying the quality of reviews to mining reviews for product or service evaluation. [10] Deepnshi sharma et al. performs their work on tourist review. Propose work has five steps. The first step is to collect the review of tourist domain. In the second step, at first, the reviews are split into sentences, then words are tagged according to their POS, next annotating the tagged word with related sentiment, and finally classifying the sentiment as positive or negative. The third step is the categories, each review into different categories like peak season, weather, expected expenditure etc. In the fourth step, review is classified as positive or negative review according to their categories. And the fifth step displays the result according to the criteria given by the user.

In [9], Cristian Bucur used hotel review from TripAdvisor website for his work for extracting opinions from website used unsupervised method and a lexical resource. System consists of two modules: a content acquisition module for collecting the reviews from website and an analysis module to preprocess the extracted data and implements opinion mining process. In analysis module, first review splits into a sentence and then tokenization process splits a sentence into component word. And lastly the polarity of word is evaluated by using SentiWordNet. N Hu et. al [12] which is used frequent item sets to extract the most relevant features from a domain and pruned it to obtain a subset of features.

They extract the nearby adjectives to a feature as an opinion word regarding that feature. Using a seed set of labeled Adjectives, which they manually develop for each domain, they further expand it using WordNet and use them to classify the extracted opinion words as positive or negative. While some researchers focus their studies on the impact of online product reviews on sales, an important question remains unanswered: so, can online product reviews reveal the true quality of the product? To test the validity of this hypothesis, B Pang and L. Lee in [7] use data from Amazon to test the underlying distribution of online reviews and try to answer this question. In summary, most of the current related work focuses on problems of opinion mining, product aspect rating, review summarization, etc. To the best of our knowledge, there has been no focused study regarding ranking products based on customer reviews.

CHAPTER 2

BACKGROUND THEORY

2.1 Data Mining

Data mining is defined as the process used to extract usable data from a larger unprocessed data set. This includes analyzing the data model in large batches of data using one or more software. Data mining has applications in several areas, such as science and research. As a data mining application, companies can learn more about their customers, develop more effective strategies related to different business functions and, therefore, use resources more optimally and insightfully. This allows companies to approach their goals and make better decisions. Data mining involves efficient data collection and storage and computer processing. Data mining uses advanced mathematical algorithms to segment the data and evaluate the probability of future events.

Data mining is also called Knowledge Discovery in Data (KDD). It is commonly defined as the process of discovering useful patterns or knowledge from data sources e.g., databases, texts, images, the Web, etc. The patterns must be valid, potentially useful, and understandable. Data mining is a multi-disciplinary field involving machine learning, statistics, databases, artificial intelligence, information retrieval, and virtualization. There are many data mining tasks. Some of the common ones are supervised machine learning (or classification), unsupervised machine learning (or clustering), association rule mining and sequential pattern mining. A data mining application usually starts with an understanding of the application domain by data analysts (data miners), who identify suitable data sources and the target data. With the data, data mining can be performed, which is usually carried out in three main steps:

- Pre-processing: The raw data is usually not suitable for mining due to various reasons. It may need to be cleaned in order to remove noises or abnormalities. The data may also be too large and/or involve many irrelevant attributes, which call for data reduction through sampling and attribute selection. Details about data pre-processing can be found in any standard data mining textbooks.

- Data mining: The processed data is then fed to a data mining algorithm which will produce patterns or knowledge.
- Post-processing: In many applications, not all discovered patterns are useful. This step identifies those useful ones for applications. Various evaluation and visualization techniques are used to make the decision.

The whole process (also called the data mining process) is almost always iterative. It usually takes many rounds to achieve final satisfactory results, which are then incorporated into real-world operational tasks [6].

Data mining is the process of finding anomalies, patterns, and correlations in large datasets to predict results. Using a wide range of techniques, it can be used to increase revenue, reduce costs, improve customer relationships, reduce risks, and more.

2.2 Opinion Mining

Opinion mining is a research subtopic of data mining that aims to automatically gain useful knowledge. It has been widely used in real-world applications such as ecommerce, business-intelligence, information monitoring, and public polls. Opinion mining attempts to determine the author's feelings, attitudes, or opinions expressed in text in relation to a particular topic. There is an increasing number of review websites on the web that allow users to post their opinions on travel agency services and provide positive or negative ratings [4]. These are important resources to advise new users and help the customer for making decision about trip. Opinions are at the heart of almost all human activities and have a significant impact on their behavior. This is true not only for individuals but also for organizations. Opinion exploration refers to the use of natural language processing (NLP) to identify and extract subjective information from online websites. Opinion mining is widely used in journals and social media for a variety of applications, from marketing to customer service.

Opinions are central to almost all human activities and are key influencers of their behaviors. The beliefs and perceptions of reality, and the choice are, to a considerable degree, conditioned upon how others see and evaluate the world. This is not only true for individual but also true for organizations. Opinions and it related

concepts such as sentiments, evaluations, attitudes and emotions are the subjects of the study of sentiment analysis and opinion mining. The inception and rapid growth of the field coincide with those of the social media on the Web, e.g., reviews, forum discussions, blogs, micro-blogs, Twitters, and social networks, a huge volume of opinionated data are collected in digital forms.

Opinions can be expressed on anything, e.g., a product, a trip, an individual, an organization, an event, a topic, etc. The components of opinions are

- Opinion Holder: Opinion holder is the person or organization that express the opinion. Opinion holders are usually the authors of the postings, although occasionally some authors cite or repeat the opinions of others. Opinion holders are more important in new articles because they often explicitly state the person or organization that hold a particular view.
- Opinion Object: It is a feature about which the opinion holder is expressing his opinion.
- Opinion orientation: Determine whether the opinion about an object is positive, negative or neutral.

Opinion mining refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. Opinion mining is widely applied to reviews and social media for a variety of applications ranging from marketing to customer service. Opinion mining can be defined as a sub-discipline of computational linguistics that focuses on extracting people's opinion from the web. The recent expansion of the web encourages users to contribute and express themselves via blogs, videos, social networking sites, etc. All these platforms provide a huge amount of valuable information that users are interested in analyzing. Given a piece of text, opinion-mining system can analyze:

- Which part is opinion expression?
- Who wrote the opinion?
- What is being commented?

Opinion mining is the part from structured data, the Web also contains a huge amount of unstructured text. Analyzing such text is also great importance. It is perhaps even more important than extracting structured data because of the sheer volume of valuable information of almost any imaginable types contained in it. The task is not only technically challenging but also very useful in practice because

business and organizations always want to know customer opinions on their products and services. For example, business always want to find public or consumer opinions on their products and services. Potential customers also want to know the opinion of existing users before they use a service or purchase a product. Moreover, opinion mining can also provide valuable information for placing advertisements in the Web pages.

2.3 Sentiment Analysis

Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. It represents a large problem space. There are also many names and slightly different tasks, e.g., sentiment analysis, opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, etc. The meaning of opinion itself is still very broad. Sentiment analysis and opinion mining mainly focuses on opinions which express or imply positive or negative sentiments.

Sentiment analysis uses natural language processing to identify and extract subjective information from online sites. Attitude holders (sources), attitude goals (aspects), and attitude types from a set of types where love, hate, value, etc. are desired. The process of extracting emotions is handled automatically. This saves time and effort. It is also one of the artificial intelligence that analyzes emotion data and has poor human emotion. Sentiment analysis is used in large-scale applications such as e-commerce, digital marketing, travel planning, and politics. Online management can be used to analyze web and social media mentions of products, services, travel agencies, marketing campaigns, or brands.

Sentiment analysis is the process of analyzing pieces written online to determine the emotional tone they carry. Sentiment analysis is used to find the client's attitude to something. Sentiment analysis tools classify written texts as positive, neutral or negative. Sentiment analysis saves time and effort because the emotion deduction process is fully automated - it is also one of the artificial intelligences that analyze emotional data and people's participation is therefore rare. Sentiment analysis and opinion polling find broad application, ranging from e-commerce, marketing,

politics and research. It can be used in online reputation management - to analyze mentions on the web and social media about a product, service, marketing campaign or brand.

Sentiment analysis is the automated process of analyzing text data and classifying opinions as negative, positive or neutral. Usually, besides identifying the opinion, these systems extract attributes of the expression e.g.:

Polarity: if the speaker expresses a positive or negative opinion,

Object: the thing that is being talked about,

Opinion holder: the person or entity expressing the opinion.

Currently, sentiment analysis is a subject of great interest and development because it has many practical applications. Businesses use sentiment analysis to automatically analyze survey responses, product reviews, comments on social media, and more.

With the explosive growth of social media (e.g., reviews, forum discussions, blogs, micro-blogs, Twitter, comments, and postings in social network sites) on the Web, individuals and organizations are increasingly using the content in these media for decision making. Nowadays, if one wants to buy a consumer product, one is no longer limited to asking one's friends and family for opinions because there are many user reviews and discussions in public forums on the Web about the product. For an organization, it may no longer be necessary to conduct surveys, opinion polls, and focus groups in order to gather public opinions because there is an abundance of such information publicly available. However, finding and monitoring opinion sites on the Web and distilling the information contained in them remains a formidable task because of the proliferation of diverse sites. Each site typically contains a huge volume of opinion text that is not always easily deciphered in long blogs and forum postings. The average human reader will have difficulty identifying relevant sites and extracting and summarizing the opinions in them. Automated sentiment analysis systems are thus needed.

2.3.1 Different Levels of Sentiment Analysis

Sentiment analysis can be applied to a range of levels.

- **Document-level** sentiment analysis gets the sentiment of the entire document or paragraph.
- **Sentence-level** emotional analysis results in a single sentence of emotion.
- **Entity and aspect level** emotion analysis results in body part expression emotions.

Document Level: This process extracts sentiment from the overall review and categorizes the entire opinion based on the overall sentiment of the opinion holder. The goal is to categorize reviews as positive, negative, or neutral. For example, "I recently visited Ngapali and am very satisfied with the hotel. The food is delicious. The sound quality of the service is excellent. I just love it!". Document-level classification works best when the document is written by one person and expresses opinions / feelings about a single entity.

Statement level: This process typically involves two steps.

- Subjectivity classification of a sentence into one of two classes: objective and subjective.
- Sentiment classification of subjective sentences into two classes: positive and negative.

Objective sentences provide some factual information, and subjective sentences represent personal feelings, views, feelings, or beliefs. Subjective sentence identification can be achieved in a variety of ways, such as the Naive Bayesian classification. However, knowing that a sentence has a positive or negative opinion is not enough. This is an intermediate step that helps to eliminate silence and to some extent determine whether the sentiment about an entity and its aspects is positive or negative. Subjective text may include multiple opinions and subjective and factual provisions.

Example "The weather is not fine. People are crowded in everywhere." Sentiment classification at both the document and sentence levels is useful, but it does not find things that people like or dislike, or target opinions.

Entity and Aspect Level: The goal of this process is to identify and extract the characteristics of the object commented by the opinion holder and determine whether the opinion is positive, negative, or neutral. Feature synonyms are grouped to create a feature-based summary of multiple reviews.

2.4 Pre-Processing of the System

Data preprocessing is essential before its actual use. Data preprocessing is the concept of changing raw data into a clean data set. The dataset is preprocessed to check missing values, noisy data, and other inconsistencies before executing it to the algorithm. Data must be in a format appropriate for Machine Learning. The pre-processing of the system includes part of speech tagging, Stanford Parser and dependency relations.

2.4.1 Part-of-Speech Tagging

A Part-Of-Speech Tagger (POS Tagger) is a piece of software that reads text in some language and assigns part of speech to each word (and other token), such as noun, verb, adjective, etc., although generally computational applications use more fine-grained POS tags like 'noun-plural'. The tagger was originally written by Kristina Toutanova. Since that time, the other researchers have improved its speed, performance, usability, and support for other language.

Part-of-speech tagging is used in this work to identify words, corresponding to part-of-speech, that are good predictors of sentiment in sentences. Part-of-speech (POS) information is commonly exploited in sentiment analysis and opinion mining. One simple reason holds for general textual analysis, not just opinion mining: part-of-speech tagging can be considered to be a crude form of word sense disambiguation.

POS tags (or part-of-speech tags) are special labels assigned to each token (word) in a text corpus, indicating part-of-speech and other grammatical categories such as tense and number (plural / singular) etc. POS tags are used in corpus search and text analysis tools and algorithms. The set of all POS tags used in the corpus is called a tag set. It can also go to a different level of detail. The basic tag set can contain only the most common part-of-speech tags (N for nouns, V for verbs, and A for adjectives). It is more common to explain in more detail and distinguish between singular and plural nouns, verb conjugations, tenses, aspects, etc.

Table 2.1: Part-of-Speech (POS) Tagging

POS Tag	Description	Example
CC	coordinating conjunction	and
CD	cardinal number	1, third

POS Tag	Description	Example
DT	determiner	the
IN	preposition, subordinating conjunction	in, of, like
JJ	adjective	green
JJR	adjective, comparative	greener
JJS	adjective, superlative	greenest
MD	modal	could, will
NN	noun, singular or mass	table
NNS	noun plural	tables
POS	possessive ending	friend's
PP	personal pronoun	I, he, it
RB	adverb	however, usually, naturally, here, good
RBR	adverb, comparative	better
RBS	adverb, superlative	best
TO	infinitive 'to'	to go
VB	verb be, base form	be
VBD	verb be, past tense	was, were

Part of speech (POS) plays an important role in identifying aspects and sentiment words. In this step, part of speech (POS) tagging is performed for each review statement. Used in reviews to identify words that are nouns, adjectives, and verbs. Nouns or noun phrases and adjectives arising from POS tags were usually expressed as aspects and sentiment words, respectively. In some cases, verbs and adverbs can also be expressed as words of emotion. The POS tags used in this work are based on the English POS tags from Penn Treebank. Table 2.1 shows the list of POS tags used to determine the dependency POS tag pattern.

2.4.2 Stanford Parser

A natural language parser is a program that determines the grammatical structure of a sentence, for example, which groups of words come together (as

"phrases") and which words are the subject or object of a verb. Probabilistic parsers attempt to generate the most likely analysis of new sentences using linguistic knowledge gained from manually parsed sentences.

This thesis paper focuses on feature extraction and opinion word extraction for customer review rankings. In the feature extraction phase, part-of-speech tagging must be performed to identify nouns / noun phrases from reviews that can be product features. Nouns and noun phrases are most likely to be product features. POS tagging is important for generating common language patterns. Use Stanford-POS Tagger to parse each sentence, generate part-of-speech tags for each word (words are nouns, adjectives, verbs, adverbs, etc.) and identify simple nouns and verb groups (syntax chunks). Opinion word extraction uses the extracted features to find the closest opinion word with an adjective / adverb. To determine the opinion orientation of each sentence, three subtasks need to be performed.

First, a series of opinion words (adjectives, usually used to express opinions) are identified. If an adjective appears near a product feature of a sentence, it is considered a word of opinion. It can extract opinion words from reviews using extracted features where both adjectives and adverbs are good indicators of subjectivity and opinion. Therefore, it needs to extract phrases that contain adjectives, adverbs, verbs, and nouns that mean opinions. It also considers some verbs (like, recommend, like, like, thank, dislike, and love) as words of opinion. Some adverbs (like not always, really, never, not whole, but absolutely, very, and well) are also considered. It collects all opinionated phrases of mostly two or three words like (adjective, noun), (adjective, noun, noun), (adverb, adjective), (adverb, adjective, noun), (verb, noun), and so forth from the processed POS-tagged review.

The resulting pattern are used to match and identify new review opinion phrases after POS tagging. However, the sentence has more likely opinion words / phrases, but is not extracted by the pattern. From these extracted patterns, most adjectives or adverbs imply an opinion on the nearest noun / noun phrase.

2.4.3 Dependency Relations

A dependency grammar describes a sentence structure as a set of dependencies. A dependency is an asymmetric binary relationship between a word called the head or governor and another word called a qualifier or dependency. Word

dependencies form a dependency tree. From this tree, the characteristics and opinions of related products can be captured by using the dependencies relations between them.

Table 2.2 Dependencies Relations

	Nominals	**Clauses**	**Modifier Words**	**Function Words**
Core Argument	nsubj obj iobj	csubj ccomp xcomp		
Non-core dependent	obj vocative expl dislocated	advcl	admov discourse	aux cop mark
Nominal dependents	nmod appos numod	acl	amod	det clf case
Coordination	**MWE**	**Loose**	**Special**	**Other**
conj cc	fixed flat compound	list parataxis	orphan goeswith reparandum	punot root dep

The admov relation is used for modifiers not only of predicates but also of others modifier words. The long form of universal dependency relation is as follows:

Table 2.3 The long form of universal dependency relation

The long form of universal dependency relation	**The long form of universal dependency relation**
acl: clausal modifier of noun (adjectival clause)	fixed: fixed multiword expression
advcl: adverbial clause modifier	flat: flat multiword expression
advmod: adverbial modifier	goeswith: goes with
amod: adjectival modifier	iobj: indirect object

The long form of universal dependency relation	**The long form of universal dependency relation**
appos: appositional modifier	list: list
aux: auxiliary	mark: marker
case: case marking	nmod: nominal modifier
cc: coordinating conjunction	nsubj: nominal subject
ccomp: clausal complement	nummod: numeric modifier
clf: classifier	obj: object
compound: compound	obl: oblique nominal
conj: conjunct	orphan: orphan
cop: copula	parataxis: parataxis
csubj: clausal subject	punct: punctuation
dep: unspecified dependency	reparandum: overridden disfluency
det: determiner	root: root
discourse: discourse element	vocative: vocative
dislocated: dislocated elements	xcomp: open clausal complement
expl: expletive	

2.5 Extracting Features

It is nowadays becoming quite common to be working with datasets of hundreds (or even thousands) of features. If the number of features becomes similar (or even bigger!) than the number of observations stored in a dataset then this can most likely lead to a Machine Learning model suffering from overfitting. In order to avoid this type of problem, it is necessary to apply either regularization or dimensionality reduction techniques (Feature Extraction). In Machine Learning, the dimensional of a dataset is equal to the number of variables used to represent it.

Feature Extraction aims to reduce the number of features in a dataset by creating new features from the existing ones (and then discarding the original features). These new reduced set of features should then be able to summarize most of the information contained in the original set of features. In this way, a summarized version of the original features can be created from a combination of the original set.

Feature extraction is a process of dimensionality reduction by which an initial set of raw data is reduced to more manageable groups for processing. A characteristic of these large data sets is a large number of variables that require a lot of computing resources to process. Feature extraction is the name for methods that select and /or combine variables into features, effectively reducing the amount of data that must be processed, while still accurately and completely describing the original data set.

The physical attributes of an object are called its features. Features are parsed as noun or noun phrases and are represented as _NN or _NNS. In general, the words those indicating most trip features are nouns or noun phrases. Therefore, the next step is to identify a noun phrase as a trip feature candidate. A linguistic filtering pattern is used to extract noun phrase. A process of dimensionality reduction by which an initial set of raw data is reduced to more manageable groups for processing. Feature extraction can also reduce the amount of redundant data for a given analysis. There are two methods to extract such features including frequency-based feature extraction and dependency grammar-based feature extraction.

2.5.1 Frequency-Based Feature Extraction

In this method, a set of nouns and noun phrases is gained per document. For this purpose, the words with part-of-speech tag of “N” are known as noun and the set of nouns with part-of-speech tag of “NN” are considered as noun phrases and will be added to set of nouns in such a document. As an example, in the sentence "university environment was extremely good", the phrase, "university environment" as a noun phrase and "environment" and "university" each as a noun are selected and added to set-of-words. At the next step, we determine the number of each of the nouns (bag-of-word) gained at the previous step among total current lists. To do this, a new set including all words extracted at the previous stage is constructed and then, the frequency of each word is specified. At the next step, nouns with a frequency higher than a threshold are extracted as important features. Frequency threshold can be any number, which is usually determined by experience. At the final step, it can be used the following idea to extract features with a frequency lower than defined frequency threshold. The opinion words can be utilized to describe different features. For example, noun phrase "university environment" in the previous example is selected as a feature and tagged in the documents; considering the sentence in the previous

example, a commenter has used the word "good" to describe this feature. Now, it can search the word "good" in entire documents and then extract the noun found before it as a feature. As a result, in a sentence like "university staff were very good". "University staff" is extracted as a feature. Opinion words are found using the general lexicon constructed at the first stage.

2.5.2 Dependency Grammar-Based Feature Extraction

The polarity of consumer sentiment for travel functions is determined by analyzing the dependencies between functional and emotional terms. Stanford Parser is used to perform functional sentiment term dependency analysis. The result of the dependency analysis is a dependency tree and dependency pairs of a set of feature and emotion terms. Each review generates a dependency tree. Next, dependencies are identified that may include both trip features and sentiment terms. The system searches for functional term dependencies and finds subject-predicate relations (nsubj), verb-object relations (dobj), adjective modification relations (amod), and relative clause modification relations (rcmod). Identify Dependency grammar refers to the structure of a set of dependencies. Dependencies indicate the relationship between a specified feature and its associated opinion.

2.6 Extracting Opinion

The next phase after features extraction is the extraction of opinionated words used on the trip features in the reviews. These are words that are primarily used to express subjective opinions. The method adopted for extracting the opinion words in this work is built on some established facts about distinguishing sentences used to express subjective opinions from sentences used to objectively describe some factual information. As a result, it can adjectives tagged by the POS tagger as opinion words and limit the opinion words extraction to those sentences that contain one or more product features, as the interest is in customers' opinions on product features.

Opinion words are usually feeling or attitudes of the writer. In this system, opinion words are extracted by using adjective words. Adjective words are represented as _JJ or _JJS. Extracting opinion words with relevant features are

processed in this phase [2]. Extracting opinion words are very important in this step so that it can be used to extract useful information in a document.

2.7 WordNet

In 1993, WordNet was introduced. It is a lexical database, organized as a semantic network. The development began in 1985 at Princeton University by a group of psychologists and linguists, and the university still is the maintainer of this lexical database. Even though it was not created with the intention to serve as knowledge source for tasks in computational linguistics, it has been used as such. It has been widely used as a lexical resource for different tasks, have been ported to several different languages, and has spawned many different subsets. One task that it has been widely used for is the previously mentioned word sense disambiguation (WSD).

The idea behind WordNet is to create a "dictionary of meaning" integrating the functions of dictionaries and thesauruses. Lexical information is not organized in word forms, but in word meanings which is consistent with a human representation of meaning and their processing in the brain.

2.7.1 Structure and Contents of WordNet

WordNet consists of four separate databases, one for nouns, one for verbs, one for adjectives and one for adverbs. It does not include closed class words. The current version available for download is WordNet 3.0, which was released in December 2006. It contains 117,097 nouns, 22,141 adjectives, 11,488 verbs and 4,601 adverbs. There is a later release, 3.1, which is available for online usage only.

The basic structure is synsets. These are sets of synonyms, or more correct, near-synonyms, since there exists none to few true synonyms. Synsets contains a set of lemmas, and these sets are tagged with the sense they represent. These senses can be said to be concepts, all of the lemmas (or words), can be said to express the same concept. Word forms which have different meanings appear in different synsets.

2.7.2 SentiWordNet 3.0

WordNet is an English word that is grouped into a set of synonyms called synsets, provides a short, general definition, and records the various semantic

relationships between these synonym sets. Nouns, verbs, adjectives, and adverbs are distinguished because they follow different grammatical rules. It does not include prepositions and determinants. It can also be used to find synonyms for all words. SentiWordNet (SWN) is an extension of WordNet that aims to increase information about verbal sentiment. In SWN, each synset has a positive score, a negative score, and an objective (neutral) score. The sum of these three scores equals 1, indicating the relative strength of the positive, negative, and objectivity of each synset. Consisting of tens of thousands of words, there are meanings, expressed parts of speech, and a range of words from 0 to 1 that are positive and negative. The same part of speech can have different meanings respectively, and also sentiwordnet was designed by ranking the subjectivity of all terms / synsets according to the part of speech to which the term belongs. The parts of speech represented by the sentiwordnet are adjective, noun, adverb and verb which are represented respectively as 'a', 'n', 'r', 'v'. The database has five columns, the part of speech, the offset which is a numerical ID, that when matched with a particular part of speech, identifies a synset: positive score, negative score (bottom from 0 to 1) and synset terms.

2.8 Term Frequency–Inverse Document Frequency

TF-IDF (term frequency-inverse document frequency) is a metric that represents how ‘important’ a word is to a document in the document set. It has many uses, most importantly in automated text analysis, and is very useful for scoring words in machine learning algorithms for Natural Language Processing (NLP).

TF-IDF was invented for document search and information retrieval. It works by increasing proportionally to the number of times a word appears in a document but is offset by the number of documents that contain the word. So, words that are common in every document, such as this, what, and if, rank low even though they may appear many times, since they don’t mean much to that document in particular.

However, if the word *Bug* appears many times in a document, while not appearing many times in others, it probably means that it’s very relevant. For example, if what we’re doing is trying to find out which topics some NPS responses belong to, the word *Bug* would probably end up being tied to the topic Reliability, since most responses containing that word would be about that topic.

2.8.1 How is TF-IDF Calculated?

TF-IDF for a word in a document is calculated by multiplying two different metrics:

- The **term frequency** of a word in a document. There are several ways of calculating this frequency, with the simplest being a raw count of instances a word appears in a document. Then, there are ways to adjust the frequency, by length of a document, or by the raw frequency of the most frequent word in a document.
- The **inverse document frequency** of the word across a set of documents. This means, how common or rare a word is in the entire document set. The closer it is to 0, the more common a word is. This metric can be calculated by taking the total number of documents, dividing it by the number of documents that contain a word, and calculating the logarithm.
- So, if the word is very common and appears in many documents, this number will approach 0. Otherwise, it will approach 1.

Multiplying these two numbers results in the TF-IDF score of a word in a document. The higher the score, the more relevant that word is in that particular document.

The **term frequency** of a word in a document is the easiest way to calculate the frequency. The frequency of a word is calculated by counting the number of occurrences of the word in the document and dividing by the total number of words in the document.

$$TF(t) = \frac{Number\ of\ times\ term\ t\ appears\ in\ a\ document}{Total\ number\ of\ terms\ in\ the\ document} \qquad 2.1$$

The **inverse document frequency** means how common or rare a word is in the entire document set. The value of a word is closer to zero, the more common a word is. This formula can be calculated by taking the total number of documents, dividing it by the number of documents that contain a word, and calculating the logarithm.

$$IDF(t) = \log_e\left(\frac{Total\ number\ of\ documents}{Number\ of\ documents\ with\ term\ t\ in\ it}\right) \quad 2.2$$

The **TF-IDF** is calculated by multiplying the results in the TF and IDF score of a word in a review document. The higher the score, that word is more relevant in that document.

$$TF - IDF(t) = TF(t) * IDF(t) \quad 2.3$$

2.9 Ranking the Features

The overall weight of each review is calculated by adding the score of the opinion word with the number of sentences as following equation (4).

$$Total\ Weight\ t = \sum_{t=1}^{n} (wt - of - positive - feature\ - wt - of - negative - feature) \quad 2.4$$

where n is the number of sentence and t is the term in this review. If the total weight of a feature is positive, then that review is termed as positive and is thought to be likely by the customer. Similarly, a negative weight indicates the feature is not liked by the user [6]. The total weight of the sentence or document is equal to the total weight of the positive features minus the total weight of the negative features.

CHAPTER 3

OPINION MINING SYSTEM OF CUSTOMER REVIEWS BY USING FEATURE EXTRACTION

With the rapid growth of the Internet, web opinion sources are emerging dynamically, and both potential customers and service providers are serving for prediction and decision-making purposes. As a result, opinion mining methods that automatically process customer mining to extract product features and user opinions expressed about them have become popular. To overcome the task of manually scanning a large number of reviews one by one, people automatically process various reviews and provide the necessary information in the right format to help both potential customers and service providers who are interested in this tourism field. By applying dependencies, this system can properly identify the semantic relationships between service features and opinions. The implemented system also used SentiWordNet to find numerical scores for all features. The system aims to collect customer reviews from the tourism sector and extract relevant features and opinions to evaluate the service.

Figure 3.1 Step by Step System Flow

Step 1: Read new review

Step 2: For each review, part-of-speech tagging and dependency relations are performed as preprocessing step.

Step 3: Trip feature candidates are extracted using Frequency-based feature extraction.

Step 4: Opinion words are extracted.

Step 5: The extracted opinion words are related with corresponding features by using dependency relation.

Step 6: Then, the sentiment orientation and score of the opinion words are identified with the help of SentiWordNet according to the Equation (4).

Step 7: Calculate the total weight of the document/review according to the total weight of these features.

Step 8: Find the specified features using Term Frequency- Inverse Document Frequency (TF-IDF) to show the reviewer.

8.1: Count the features in a review.

8.2: Sum the total counted times of features.

8.3: Calculate the Term Frequency using Equation (1).

8.4: Calculate the Inverse Document Frequency using Equation (2).

8.5: Calculate Term Frequency- Inverse Document Frequency (TF-IDF) using Equation (3).

3.1 Preprocessing

Preprocessing is essential before its actual use. In data mining, preprocessing refers to the data preparation and transformation steps that are applied to raw data before it is fed into a data mining algorithm or model. The primary objective of data preprocessing is to enhance the quality of the data and make it suitable for analysis. It involves several techniques to clean, integrate, transform, and reduce the data.

For each review, part-of-speech tagging, and dependency relations are performed as preprocessing step.

3.1.1 POS Tagging using Stanford Parser

Stanford Parser is a natural language processing tool that provides various linguistic analysis capabilities, including Part-of-Speech (POS) tagging. POS tagging is the process of assigning grammatical labels or tags to each word in a sentence, indicating its syntactic category and role in the sentence's structure. These tags help in understanding the grammatical relationships between words and their context, which is essential for many language processing tasks like parsing, information extraction, and sentiment analysis. At first, input of a trip review can be seen as follows:

This is the review for “Myanmar Tour Asia Agency”.

Figure 3.2 Sample Review for "Myanmar Tour Asia Agency"

I just got back from my wonderful trip to Bagan and I would like to thank Myanmar Tour Asia Agency for making sure that I got a memorable trip I have ever been. The staffs were patient and informative whenever I had some inquiries about the service or the trip or the food. All of the staffs knew exactly how to treat the guests with love and care. In fact, I love the food that they served during the trip. (My mouth is now watering just by the thought of that.) Overall, I would certainly recommend everyone to go on a trip to Bagan with Myanmar Tour Asia Agency.

For each review, part-of-speech tagging and dependency relations are performed at the preprocessing step. POS tagging usually works by specifying the character string that if matched. A POS tagging helps to decide which character string is noun/nouns (NN or NNS), adjectives (JJ/JJR/JJS), adverbs (RB/RBR/RBS), verbs (VB/VBD/VBN), determiner (DT), preposition (PP), infinitive to (TO), personal pronoun (PP) and so on. In this phase, trip review is parsed using Stanford Parser to generate the POS tagsets and dependency relations.

Table 3.1 POS Tagging

Word/POS-Tag Name	**Word/POS-Tag Name**
I/PRP	ever/RB
just/RB	been/VBN
got/VBD	**./.**
back/RP	The/DT
from/IN	staffs/NNS
my/PRP	were/VBD
wonderful/JJ	patient/JJ
trip/NN	and/CC
to/TO	informative/JJ
Bagan/NNP	whenever/WRB
and/CC	I/PRP
I/PRP	had/VBD
would/MD	some/DT
like/VB	inquiries/NNS

Word/POS-Tag Name	Word/POS-Tag Name
to/TO	about/IN
thank/VB	the/DT
Myanmar/NNP	service/NN
Tour/NNP	or/CC
Asia/NNP	the/DT
Agency/NNP	trip/NN
for/IN	or/CC
making/VBG	the/DT
sure/JJ	food/NN
that/IN	./.
I/PRP	.
got/VBD	.
a/DT	.
memorable/JJ	.
trip/NN	.
I/PRP	.
have/VBP	.

3.1.2 Dependency Relations or Universal dependencies

A dependency relationship shows the relationship between the specified features and its related opinion words. The following diagram shows the dependency relationship of a trip review. A review sentence is spited to form a dependency tree in the POS tagging step using Stanford Parser. After parsing the review, the sentence is transformed into dependency relations as follows:

Table 3.2 Universal Dependency Relations

Opinion and Feature Relationship	Opinion and Feature Relationship
nsubj(got-3, I-1)	compound(Agency-20,Tour-18)
advmod(got-3, just-2)	compound(Agency-20,Asia-19)
root(ROOT-0, got-3)	dobj(thank-16, Agency-20)
compound:prt(got-3, back-4)	mark(making-22, for-21)

Opinion and Feature Relationship	Opinion and Feature Relationship
case(trip-8, from-5)	acl(Agency-20,making-22)
nmod:poss(trip-8, my-6)	xcomp(making-22, sure-23)
amod(trip-8, wonderful-7)	mark(got-26, that-24)
nmod(got-3, trip-8)	nsubj(got-26, I-25)
case(Bagan-10, to-9	ccomp(making-22, got-26)
nmod(got-3, Bagan-10)	det(trip-29, a-27)
cc(got-3, and-11)	amod(trip-29, memorable-28)
nsubj(like-14, I-12)	dobj(got-26, trip-29)
aux(like-14, would-13)	nsubj(been-33, I-30)
conj(got-3, like-14)	aux(been-33, have-31)
mark(thank-16, to-15)	advmod(been-33, ever-32)
xcomp(like-14, thank-16)	acl:relcl(trip-29, been-33)
compound(Agency-20, Myanmar-17)	…….

3.2 Extracting Features

Extracting features is the main process of the system because a lot of features are generated after parsing the part-of-speech tagging at the pre-processing step. Mostly, features are noun or noun phrases and are represented as _NN or _NNS. Therefore, the system collects the noun or noun phrases such as trip features.

Table 3.3 Extracting Features

Statements	Features (NN/NNS) Nouns
1	Trip
2	Staff, Inquiries, Services, Trip, Food
3	Staffs, Guests, Love, Care
4	Fact, Food, Trip
5	Mouth, Thought
6	Everyone, Trip

The above table shows the trip features from the trip review sentence. There are a lot of trip features. Therefore, the system needs to generate the frequently trip

features that are frequently mentioned by the customer. Then, infrequent trip features are removed because they are not necessary. Frequent trip feature candidates are extracted by using frequency-based feature extraction.

3.2.1 Extracting Features using Frequency-Based Feature Extraction

Frequency-based feature extraction extracts the frequent trip features that are frequently used in that review sentence. Moreover, frequency-based feature extraction counts the frequent trip features from the review sentence as follows:

Table 3.4 Extracting features using Frequency-Based Feature Extraction

Feature	Feature Count
Trip	5
Staff	2
Inquiries	1
Services	1
Food	2
Guests	1
Love	1
Care	1
Fact	1
Mouth	1
Thought	1
Everyone	1
Total	18

3.3 Extracting Opinion

Extracting opinion is the next process after extracting features. This step extracts the opinion words from the trip review sentences. Opinion words are extracted with three opinion types such as verb (VB: base form, VBD: past tense, VBN: past participle, VBG: gerund), adverb (RB: adverb, RBR: comparative adverb, RBS: superlative adverb) and adjective (JJ: adjective, JJR: comparative adjective, JJS:

superlative adjective). The following table shows the opinion words from the review sentences.

Table 3.5 Extracting Opinion

Statements	Opinion /JJ/JJR/JJS Adjectives	RB/RBR/RBS Adverbs	VB/VBD/VBN Verbs
1	Wonderful, Sure, Memorable	Just, Ever	Got, Thank, Like, Making, Got, Have, Been
2	Patient, Informative		Were, Had
3		exactly	Knew, Treat
4			Love, Served
5		Now, Just	Are, Watering
6		Certainly	Recommend, Go

3.4 Extracting Dependency Pairs (nsubj/dobj/amod/rcmod) based on POS

The extracted opinion words are related with corresponding features by using dependency relation. This step extracts the dependency pairs by using the four types of dependency pairs. These four types of dependency pairs are the pairs of nsubj, dobj, amod, and rcmond namely subject-predicate relationships (nsubj), verb-object relationships (dobj), adjectival modifying relations (amod), and relative clause modifying relations (rcmod). A dependency pair shows the relationship between the specified features and its related opinion words. The below table is the dependency pairs for the trip review.

Table 3.6 Extracting Dependency Pairs (nsubj/dobj/amod/rcmod) based on POS

Statements	Dependency Pairs
1	nsubj(got-3, I-1)nsubj(like-14, I-12)nsubj(got-26, I-25)nsubj(been-33, I-30) dobj(thank-16, Agency-20)dobj(got-26, trip-29) amod(trip-8, wonderful-7)amod(trip-29, memorable-28)
2	nsubj(patient-4, staffs-2)nsubj(informative-6, staffs-2)nsubj(had-9, I-8) dobj(had-9, inquiries-11)dobj(had-9, trip-17)

Statements	Dependency Pairs
3	nsubj(knew-5, All-1) dobj(treat-9, guests-11)
4	nsubj(love-5, I-4)nsubj(served-10, they-9) dobj(love-5, food-7)
5	nsubj(watering-6, mouth-3)
6	nsubj(recommend-6, I-3) dobj(recommend-6, everyone-7)

The above table shows many dependency pairs so that the system needs to produce more relevant dependency pairs. By using dependency grammar-based feature extraction, the system can get the most useful trip information.

3.4.1 Dependency Grammar-Based Feature Extraction

Dependency grammar-based feature extraction means the dependency pairs which are related with the extracted features. So, the system can produce the relationship between the extracted features and related opinion words to show the important trip review data.

Table 3.7 Dependency Grammar-Based Feature Extraction

Statements	Extracting Features and Opinion
1	dobj(got-26, trip-29)amod(trip-8, wonderful-7)amod(trip-29, memorable-28)
2	nsubj(patient-4, staffs-2)nsubj(informative-6, staffs-2)dobj(had-9, inquiries-11)dobj(had-9, trip-17)
3	dobj(treat-9, guests-11)
4	dobj(love-5, food-7)
5	nsubj(watering-6, mouth-3)
6	dobj(recommend-6, everyone-7)

3.5 SentiWordNet 3.0

This is the work process for SentiWordNet 3.0. SentiWordNet 3.0 has five steps: POS, offset, posScore (positive score), negScore (negative score), and synsetterms. The parts of speech represented in SentiWordNet are adjectives, nouns, adverbs, and verbs, represented by "a", "n", "r", and "v", respectively. The database has five columns of parts of speech, offsets that are numeric IDs that identify the synth when it matches a particular part of speech, positive scores, negative scores (from 0 to 1) and synset terms.

Table 3.8 SentiWordNet 3.0

Fields	**Descriptions**
POS	Part of Speech linked with synset. This can take four possible values: a=adjective (JJ) n=noun (NN) v=verb (VBN) r=adverb (RB)
Offset	Numerical ID which associated with part of speech uniquely identifies a synset in the database.
PosScore	Positive score for this synset. This is a numerical value ranging from 0 to 1.
NegScore	Negative score for this synset. This is a numerical value ranging from 0 to 1.
SynsetTerms	List of all terms included in this synset.

Then, the sentiment orientation and score of the opinion words are identified with the help of SentiWordNet 3.0. The positive score and negative score of each synset term is shown in below (range is between 0 and 1). The total weight of trip review can get by calculating the total weight of all positive score minus the total weight of all negative score. This is the final score of each review. After checking the final weight of each trip review, it is calculated. If the final weight of each trip review is positive, that travel agency is liked by the customer and if it is negative, customer dislike this travel agency and trip.

Table 3.9 Posscore and Negscore of Each Synset

POS	Offset	PosScore	NegScore	SynsetTerms
v	02359340	0	0	got#2
a	01676517	0.75	0	wonderful#1
a	00399533	0.25	0.125	memorable#1
n	10405694	0.125	0	patient#1
n	01304570	0.125	0	informative#3
v	02740745	0	0.25	had#10
v	00078760	0	0.125	treat#3
v	01465668	0.5	0	love#1
v	00278403	0	0	watering#2
v	00882948	0.5	0	recommend#2
Total Weight		2.25	0.5	

3.6 Calculate the Final Weight of Each Trip Review

This is the formula of the final weight of each trip review. The total weight of trip review can get by calculating the total weight of all positive score minus the total weight of all negative score. This is the final score of each review.

$$Wt = \sum_{t=1}^{n} (Weight\ of\ Positive\ features - Weight\ of\ Negative\ features) \quad 3.1$$

Total Weight of sample review = (0-0)+(0.75-0)+(0.25-0.125)+(0.125-0)+(0.125-0)+
(0-0.25)+(0-0.125)+(0.5-0)+(0-0)+(0.5-0)
= 0+0.75+0.125+0.125+0.125-0.25-0.125+0.5+0.5
= 1.75(positive)

After this step is finished, the step by step calculation of each trip review is finished. But in the system implementation, the system has to show the trip features which are frequently mentioned by the customer. In section 3.2, there are many trip features so that the system cannot show the customer all these features. To select the five specified trip features, the system added the term frequency–inverse document frequency.

3.7 Term Frequency–Inverse Document Frequency

TF-IDF (term frequency-reverse document frequency) is a measure of the importance of a word to the documents in a document set. The TF-IDF of a word in a document is calculated by multiplying two different metrics, IF and IDF.

3.7.1 Term Frequency

Term frequency of a word in a review document counts the number of times a word appears in a document.

$$TF(t) = \frac{Number\ of\ times\ term\ t\ appears\ in\ a\ document}{Total\ number\ of\ terms\ in\ the\ document} \quad 3.2$$

This is the formula of term frequency to calculate the number of times of a word in a document.

3.7.2 Inverse Document Frequency

Inverse document frequency of a word across a set of documents means how common or rare words are in the entire document set. The closer to 0 the word is more common. Inverse document frequency can be calculated using the following formula:

$$IDF(t) = \log_e\left(\frac{Total\ number\ of\ documents}{Number\ of\ documents\ with\ term\ t\ in\ it}\right) \quad 3.3$$

Moreover, the system needs to calculate inverse document frequency to show how many times of a word is appears in the entire review documents. Therefore, many review documents are needed in the system to calculate the IDF. Therefore, the system can generate the useful information for the customer. There are three review documents of the travel agency namely Myanmar Tour Asia Agency.

Document 1 –Myanmar Tour Asia Agency

I just got back from my wonderful trip to Bagan and I would like to thank Myanmar Tour Asia Agency for making sure that I got a memorable trip I have ever been. The staffs were patient and informative whenever I had some inquiries about the service or the trip or the food. All of the staffs knew exactly how to treat the guests with love and care. In fact, I love the food that they served during the trip. (My mouth are now

watering just by the thought of that.) Overall, I would certainly recommend everyone to go on a trip to Bagan with Myanmar Tour Asia Agency.

Document 2 –Myanmar Tour Asia Agency

If my friends asked me what agency they should go to concerning with travel within local areas, I would recommend Myanmar Tour Asia Agency for the hotels they provide to the guests are outstanding. The rooms of the hotels are not only clean, but also have a high standard. The room service was excellent as well. However, even though the services provided by the Myanmar Tour Asia Agency was satisfactory, a few of the employees had no idea that they should treat the guests with respect. Apart from that, everything went well on my last local trip.

Document 3 –Myanmar Tour Asia Agency

Myanmar Tour Asia Agency is a terrible agency and I would never recommend anyone to go on any trips with this agency. The services as well as the food that it offers to the guests were horrible that no one seemed satisfied and a lot of people got sick after consuming the food. The Myanmar Tour Asia Agency has mentioned in their advertisement that its goal is to provide good services and delicious food to its guests, and unfortunately as a guest, I would say that they have done the opposite. Agency is very bad!

Firstly, the system counts the frequently features that are mentioned by the customers. Then, the system sums the total number of features.

Table 3.10 Three Documents' POS Tagging

Document 1	**Document 2**	**Document 3**
Trip =5	Friends =1	Agency =3
Staff =2	Agency =1	Anyone =1
Inquiries =1	Travel =1	Trip =1
Services =1	Areas =1	Service =2
Food =2	Hotels =2	Food =3
Guests =1	Guests =2	Guests =3
Love =1	Rooms =2	One =1
Care =1	Standard =1	Lot =1

Document 1	Document 2	Document 3
Fact =1 Mouth =1 Though =1 Everyone =1	Service =2 Employees =1 Idea =1 Respect =1 Everything =1 Trip =1	People =1 Advertisement =1 Goal =1
Total=18	Total=18	Total=18

Next, calculates the TF(t) and then calculates the IDF(t) by using the above two formulas.

Table 3.11 Term Frequency and Inverse Document Frequency

Document 1
TF('Trip',Document1) = 5/18, IDF('Trip')=log(3/3) = 0
TF('Staff',Document1) = 2/18, IDF('Staff')=log(3/1) = 0.48
TF('Inquires',Document1) = 1/18, IDF('Inquires')=log(3/1) = 0.48
TF('Services',Document1) = 1/18, IDF('Services')=log(3/3) = 0
TF('Food',Document1) = 2/18, IDF('Food')=log(3/2) = 0.18
TF('Guests',Document1) = 1/18, IDF('Guests')=log(3/3) = 0
TF('Love',Document1) = 1/18, IDF('Love')=log(3/1) = 0.48
TF('Care',Document1) = 1/18, IDF('Care')=log(3/1) = 0.48
TF('Fact',Document1) = 1/18, IDF('Fact')=log(3/1) = 0.48
TF('Mouth',Document1) = 1/18, IDF('Mouth')=log(3/1) = 0.48
TF('Though',Document1) = 1/18, IDF('Though')=log(3/1) = 0.48
TF('Everyone',Document1) = 1/18, IDF('Everyone')=log(3/1) = 0.48

The system calculates TF_IDF by multiplying TF and IFD values. Multiplying these two numbers results in the TF-IDF score of a word in a document.

Table 3.12 TF_IDF by multiplying TF and IFD values

Document 1
TF-IDF('Trip',Document1)= 5/18*0=0
TF-IDF('Staff',Document1)= 2/18*0.48=0.05
TF-IDF('Inquires',Document1)= 1/18*0.48=0.03
TF-IDF('Services',Document1)= 1/18*0=0
TF-IDF('Food',Document1)= 2/18*0.18=0.02
TF-IDF('Guests',Document1)= 1/18*0=0
TF-IDF('Love',Document1)= 1/18*0.48=0.03
TF-IDF('Care',Document1)= 1/18*0.48=0.03
TF-IDF('Fact',Document1)= 1/18*0.48=0.03
TF-IDF('Mouth',Document1)= 1/18*0.48=0.03
TF-IDF('Though',Document1)= 1/18*0.48=0.03
TF-IDF('Everyone',Document1)= 1/18*0.48=0.03

After that the system chooses the specified features where the number of times term t appears in a document is greater than one or IF_IDF value is 0. By using these two condition, the system chooses the relevant specified features in a document.

Trip, Staff, Services, Food and Guests are the five specified trip features to show the customer who like or dislike in travel agency. Other reviews are calculated as like this.

CHAPTER 4

IMPLEMENTATION OF THE SYSTEM

This system is implemented using Java language to write coding and MySQL to store database about the system information. In this system, there are many steps such as preprocessing, feature extraction, opinion extraction and so on. All necessary information and resources are used in this system to describe the opinion mining system of customer reviews in tourism field. Then, this system generates the useful information for new customers for their next trip plans.

It is described that the implementation of the system with fully screenshots and system design of the whole system. This system used the Window Builder plugin to design the JFrames to attract the reviewers and viewers. Therefore, this system has beautiful graphical user interface design and this system is easy to use.

It is demonstrated the whole system by using the system flowchart of the system. This chart expresses the step by step process of the system with clearly and completely.

It is also included class diagram to show the structure of the system. A class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among objects.

Moreover, sequence diagrams are used in this chapter. A sequence diagram shows object interactions arranged in time sequence. It depicts the objects and classes involved in the scenario and the sequence of messages exchanged between the objects needed to carry out the functionality of the system. There are three sequence diagrams in this chapter. They are pre-processing, extracting features and extracting opinion.

Entity Relationship Diagram is used to express the relationship between entities and its attributes. Entity Relationship Diagram, also known as ERD, ER Diagram or ER model, is a type of structural diagram for use in database design. This

chapter consists of table designs the system and this table designs show the variable names and its data types.

4.1 System Flowchart of the System

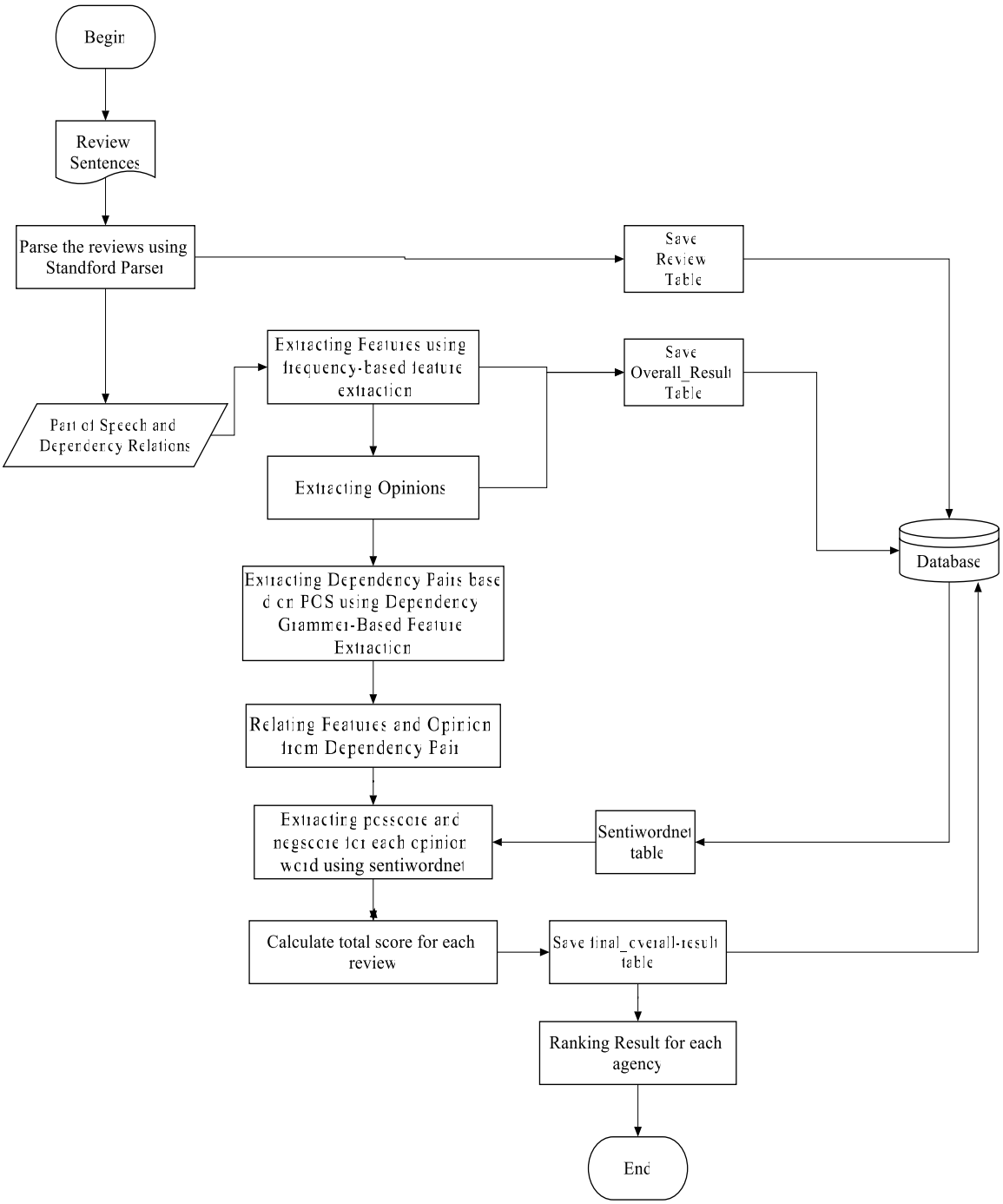


Figure 4.1 System Flowchart

4.2 Class Diagram of the System

Class diagrams are one of the most useful types of diagrams and they clearly show the structures of a particular system by modeling its classes, attributes, operations, and relationships between objects. Class diagram offers a number of benefits for any organization.

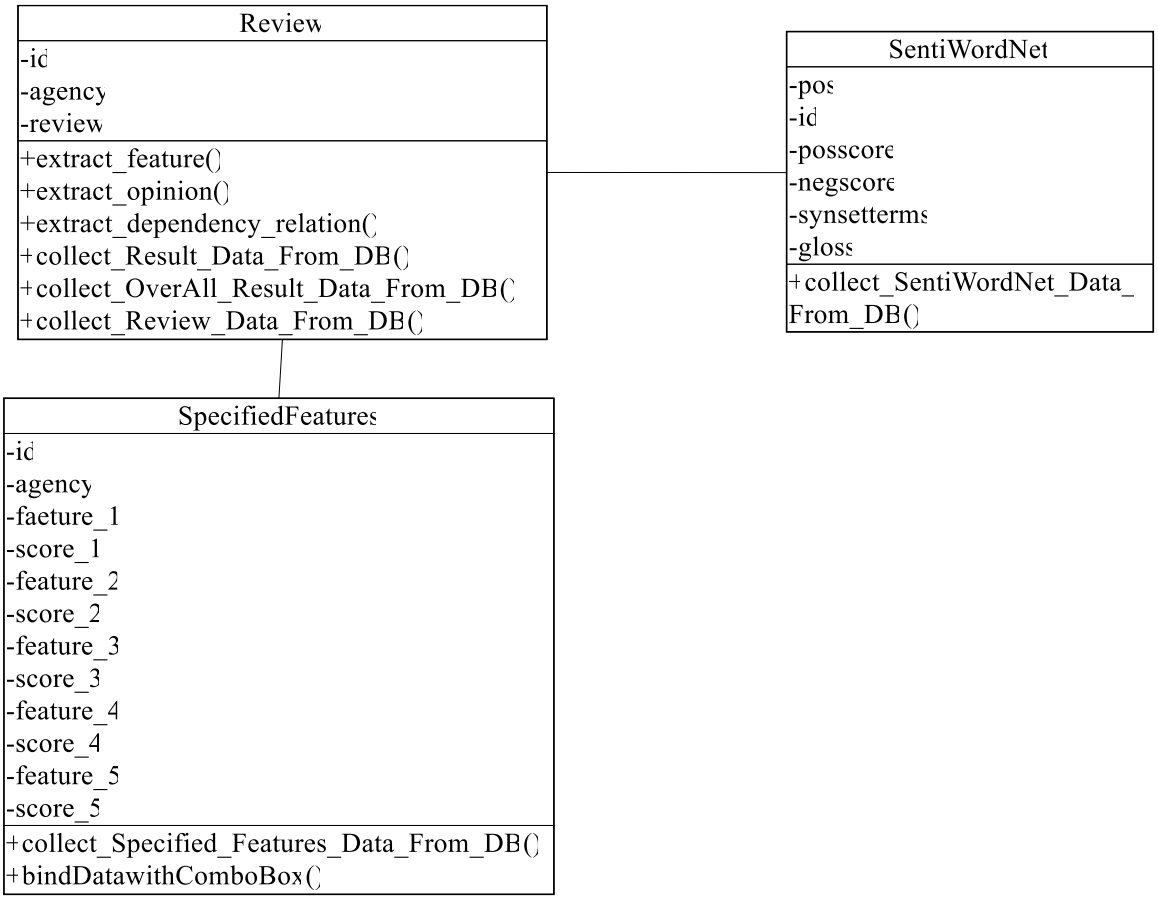


Figure 4.2 Class Diagram for Review, SentiWordNet and SpecifiedFeatures

Review class, SentiWordNet class and Specifiedfeatures class are composed of various functions. First, the system collects the reviews from the reviewer by using collect_Reciew_Data_From_DB() function and saves this reviews into the Review table. At the preprocessing step, the system extracts the review using extreact_feature() function, extract_opinion() function and extract_dependency_realtion() function. After calculating the final score of each review, the system uses the collect_Result_Data_From_DB() function and if the system wants to calculates the overall result of each agency, the system have to use collect_Overall_Result_Data_From_DB() function. Accoding to the five specified

features, the system uses the collect_Specified_Features_Data_From_DB() function. Moreover, if the reviewer wants to search each travel agency, the system shows each travel agency's most popular specified features by using bindDatawithComboBox() function. To calculate the posscore and negscore of an opinion word, the system uses collect_SentiWordNet_Data_From_DB() function.

4.3 Sequence Diagrams of Pre-Processing

Reviews are splitted into sentences. Then, these sentences are extracted to generated features and related opinion words and then the relationship between extracted features and opinion words called dependency relations by using Stanford Parser.

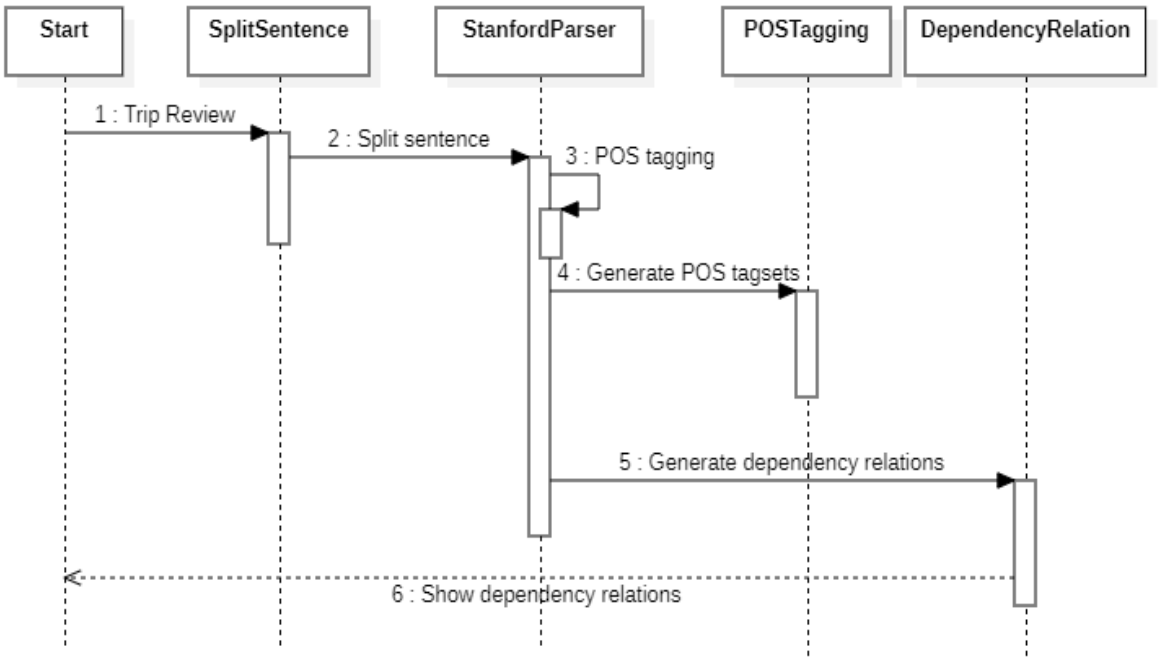


Figure 4.3 Sequence Diagram for Preprocessing

4.3.1 Sequence Diagram of Extracting Features

Features are extracted by using POS tagging which are nouns or noun phrases (_NN or NNS).

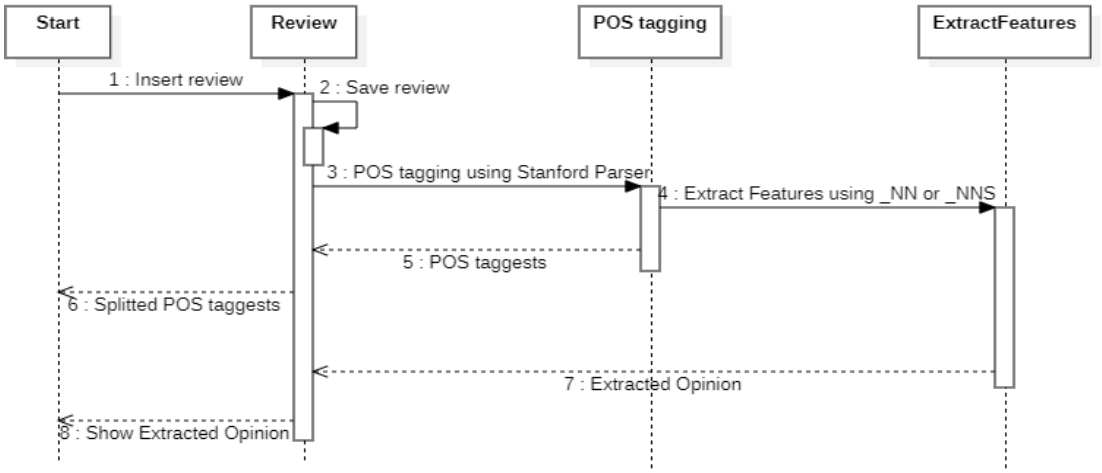


Figure 4.4 Sequence Diagram for Feature Extraction

4.3.2 Sequence Diagram of Extracting Opinion

Opinion words are extracted using POS tagging which are verb (VB: base form, VBD: past tense, VBN: past participle, VBG: gerund), adverb (RB: adverb, RBR: comparative adverb, RBS: superlative adverb) and adjective (JJ: adjective, JJR: comparative adjective, JJS: superlative adjective).

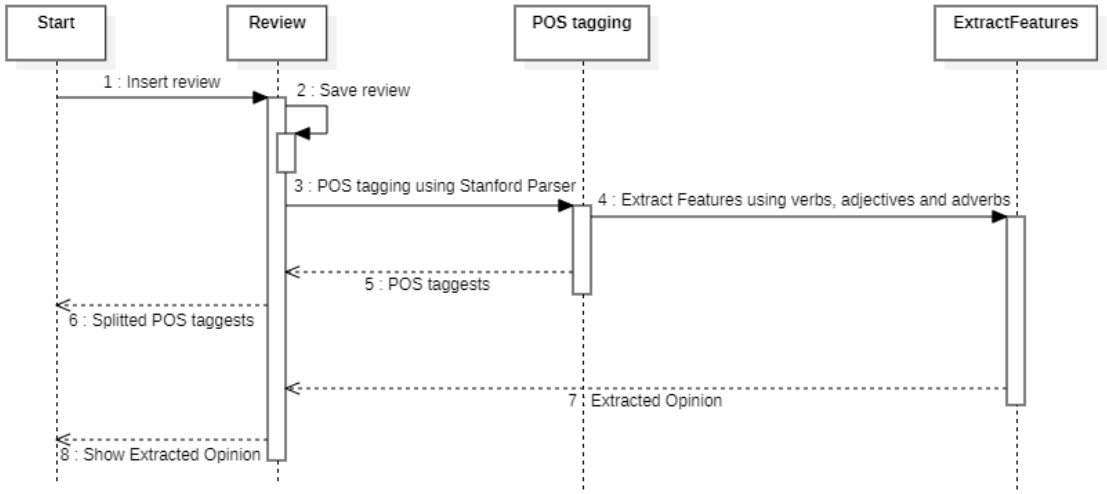


Figure 4.5 Sequence Diagram for Extracting Opinion Words

4.4 Entity Relationship Diagram of the System

ERD stands for Entity Relationship Diagram shows the relationship of entity sets stored in a database. An entity consists of attributes and its properties.

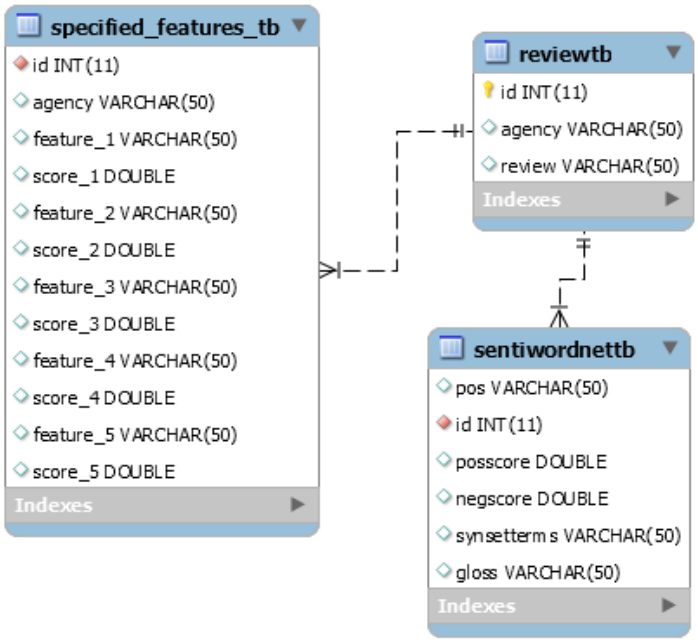


Figure 4.6 Entity Relationship Diagram for the System

4.5 Table Design View of the System

Before implementing the system, the design of each table should be described.

Table 4.1 Design View of Review Table

Review	
Id	int
Agency	varchar
Review	varchar

Table 4.2 Design View of SentiWordNet

SentiWordNet	
Pos	varchar
Id	int
Posscore	double

Negscore	double
Synsetterms	varchar
Gloss	varchar

Table 4.3 Design View of Specified_Features Table

Specified_Features	
Id	int
Agency	varchar
Feature_1	varchar
Score_1	double
Feature_2	varchar
Score_2	double
Feature_3	varchar
Score_3	double
Feature_4	varchar
Score_4	double
Feature_5	varchar
Score_5	double

4.6 Implementation of the System

This system is implemented by using Window Builder JFrame in Eclipse and MySql database to efficient and effective. The step by step testing of the system can be seen in the following section.

4.6.1 Starting of the System

When this system starts, the main form can be seen in Figure 4.10. This is the main interface of the system. In the main interface, there are four menus and they are about us, about system, objectives and welcome.

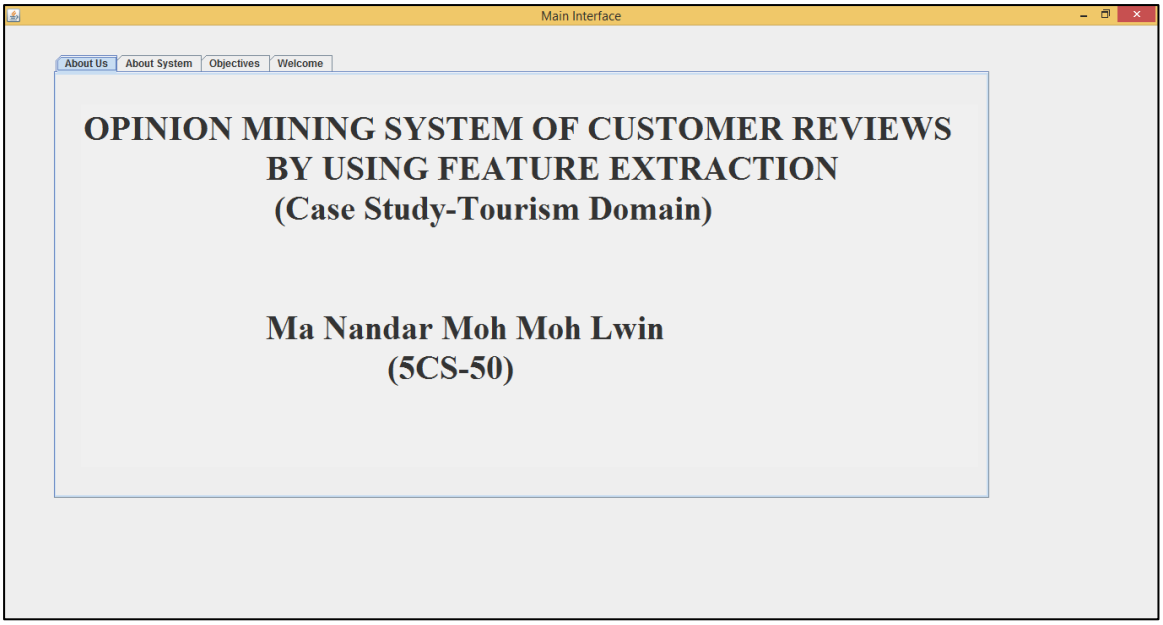


Figure 4. 7 Main Interface of the System

Firstly, the system shows the "OPINION MINING SYSTEM OF REVIEWER REVIEWS BY USING FEATURES EXTRACTION".

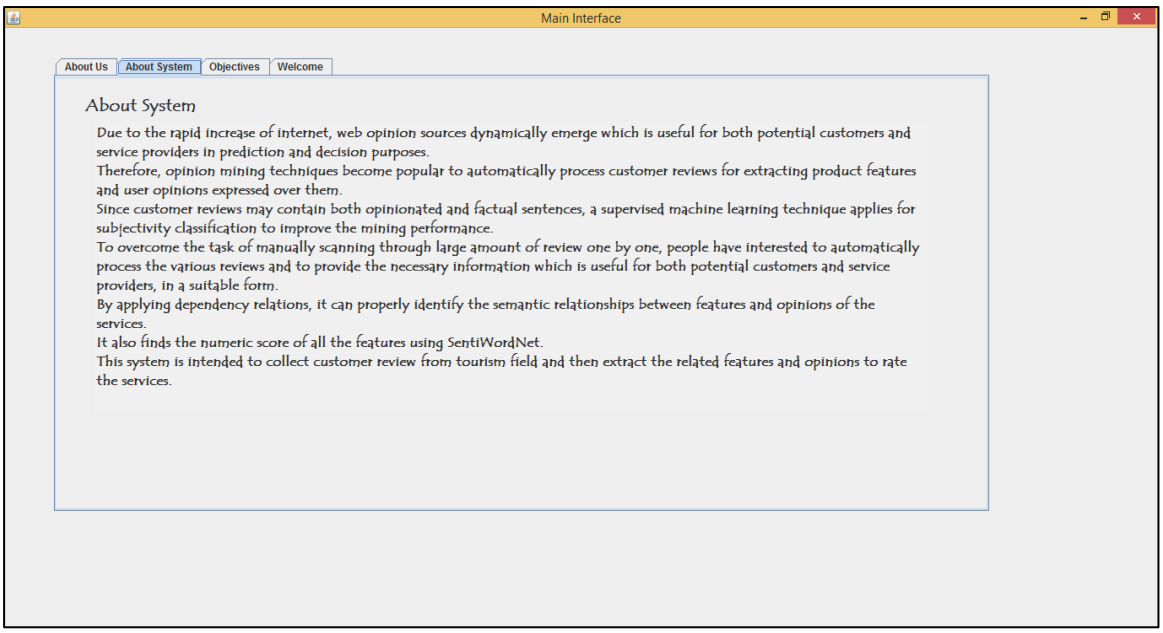


Figure 4.8 About the System

In this "About System" menu, the whole system abstract have been described to understand the reviewer why the system is created or implemented.

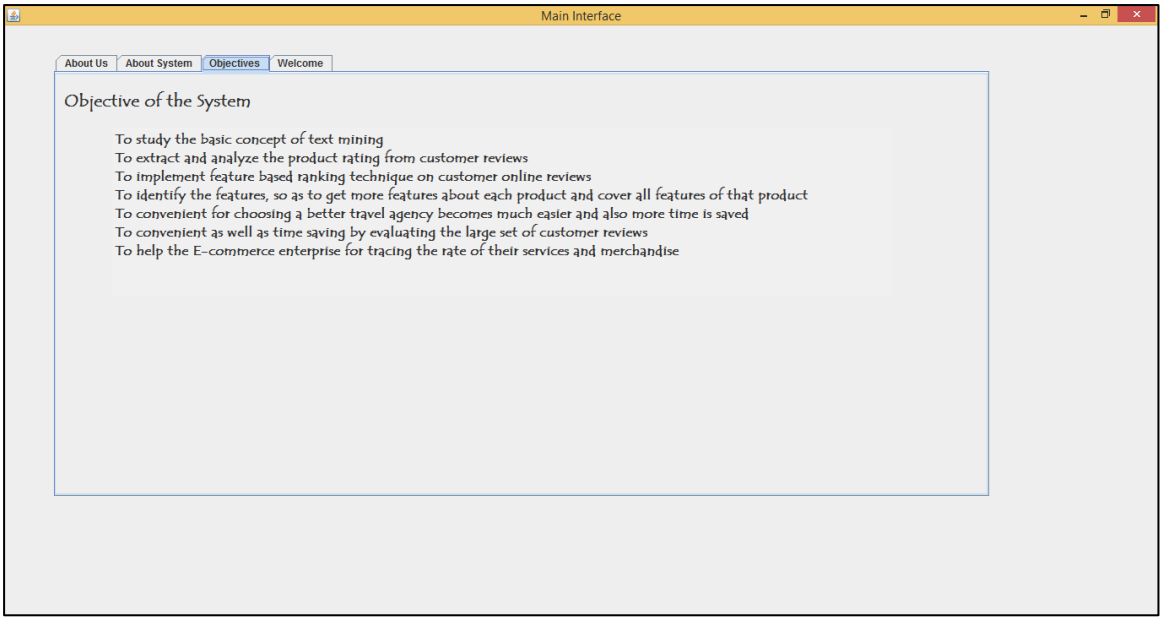


Figure 4. 9 Objective of the System

There are many objectives and they have been shown in objective menu. The reviewer can see these objectives and they can understand why the system is important in the daily routine.

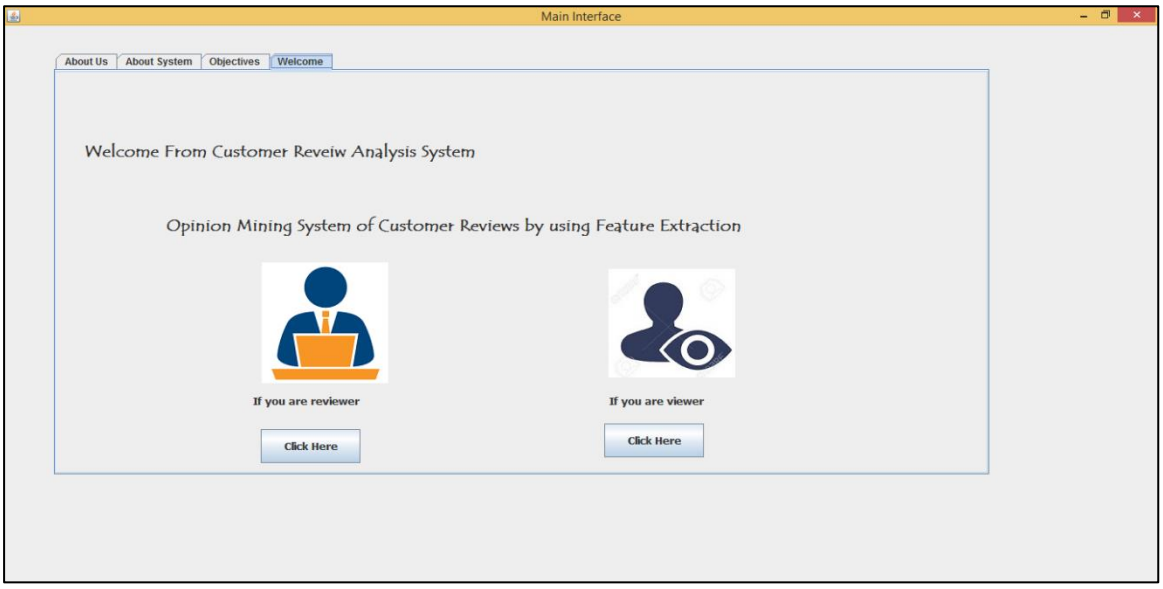


Figure 4. 10 Welcome Menu of the System

In this menu, there are two user levels and they are reviewer and viewer. By clicking the button below the review, the system will show the main interface for reviewer and by clicking the button below the viewer, the system will show the main interface for viewer.

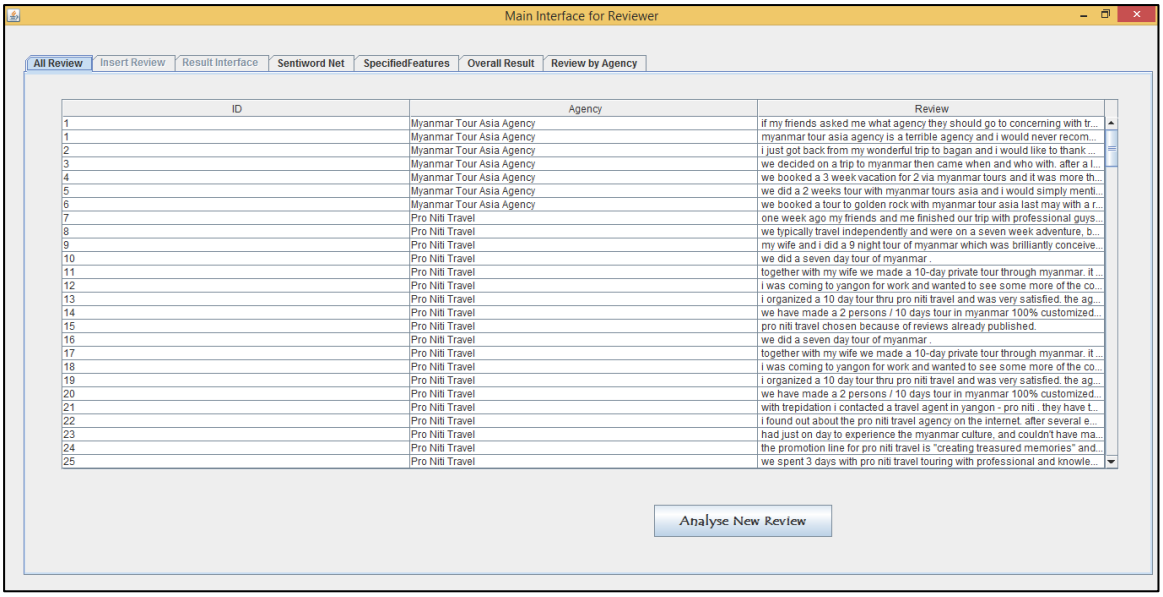


Figure 4. 11 Main Interface for Reviewer

This is the main interface for the reviewer. In this reviewer main interface, there are seven menus such as All Review, Insert Review, Result Interface, SentiWordNet, Specified Features, Overall Result and Review by Agency. All reviews that are posted by the reviewers have been shown with id number, agency name and review sentence in All Review menu.

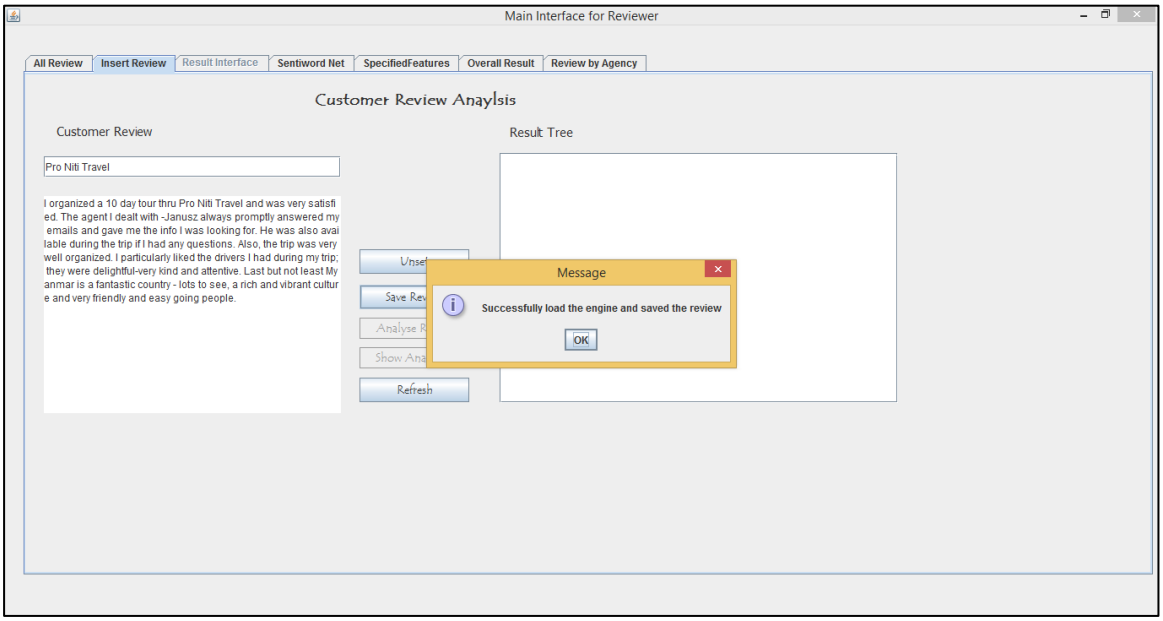


Figure 4. 12 Insert Review

If the reviewer wants to post a trip review into the system, click the Insert Review menu in the tabbed pane. Reviewer can insert the agency that he wants to review and a trip review document.

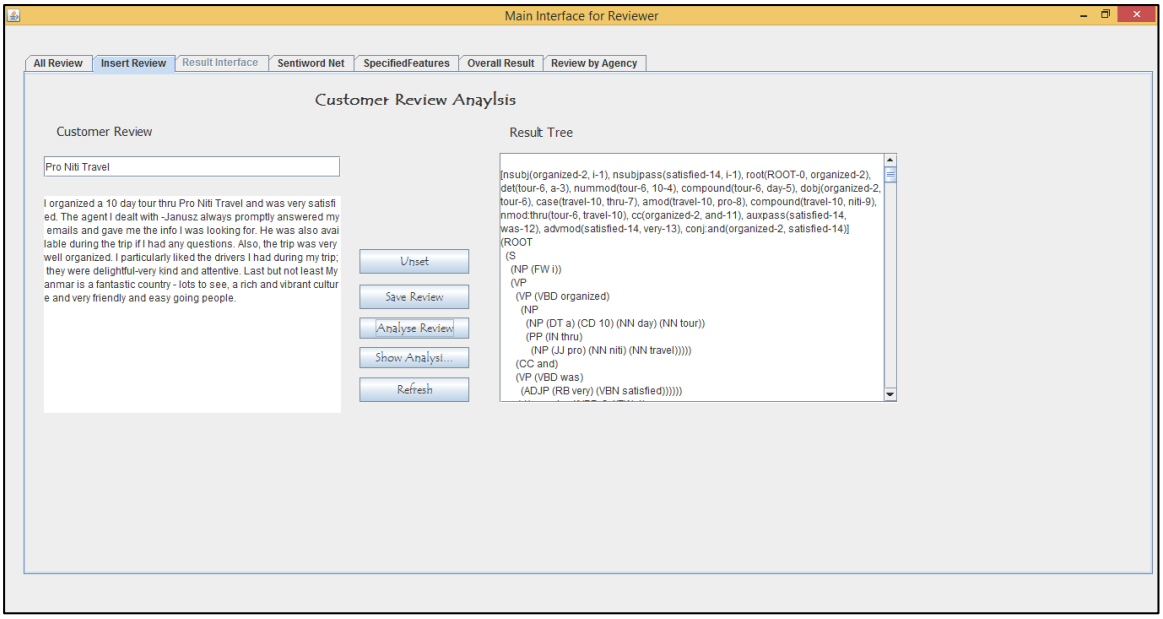


Figure 4. 13 Analyse Review using Stanford Parser

After inserting the review, the system saves the review into the review table and analyses the review using Stanford Parser as shown in Figure 4.15.

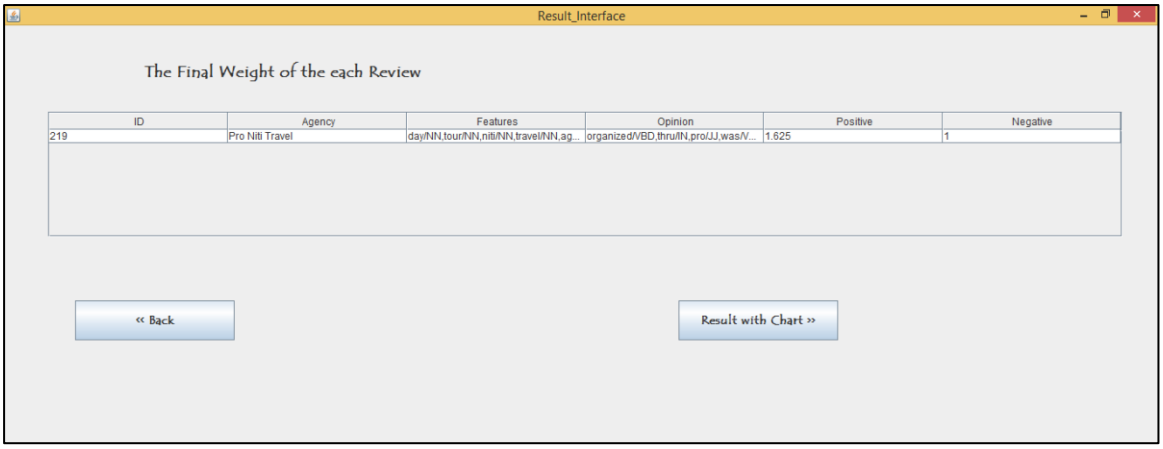
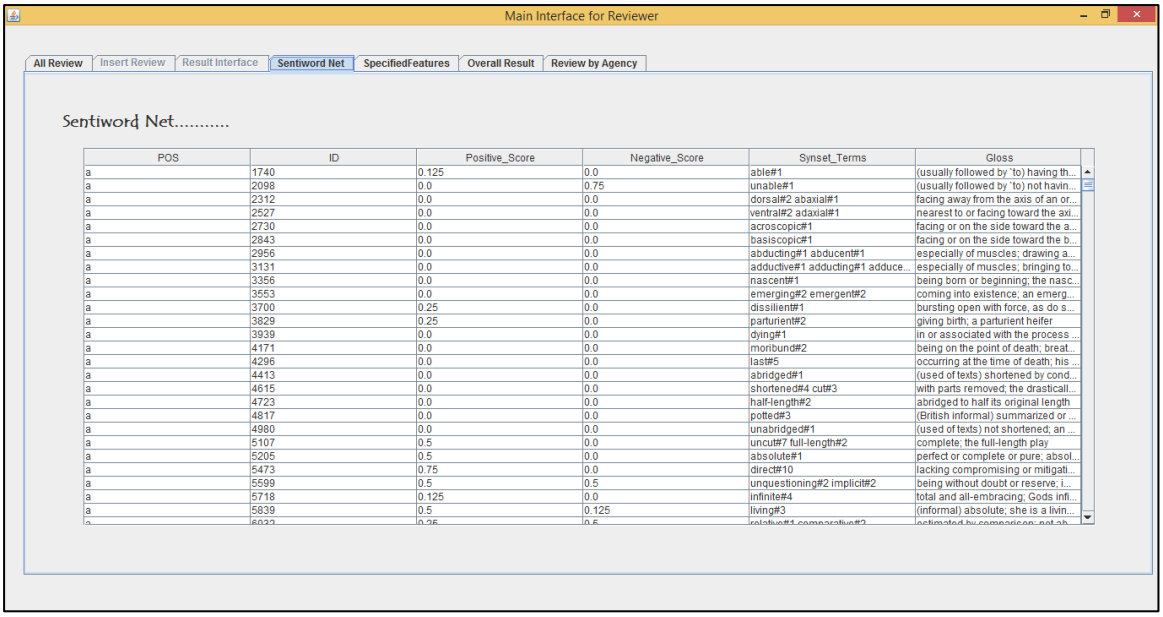


Figure 4. 14 Final Weight of a Trip Review

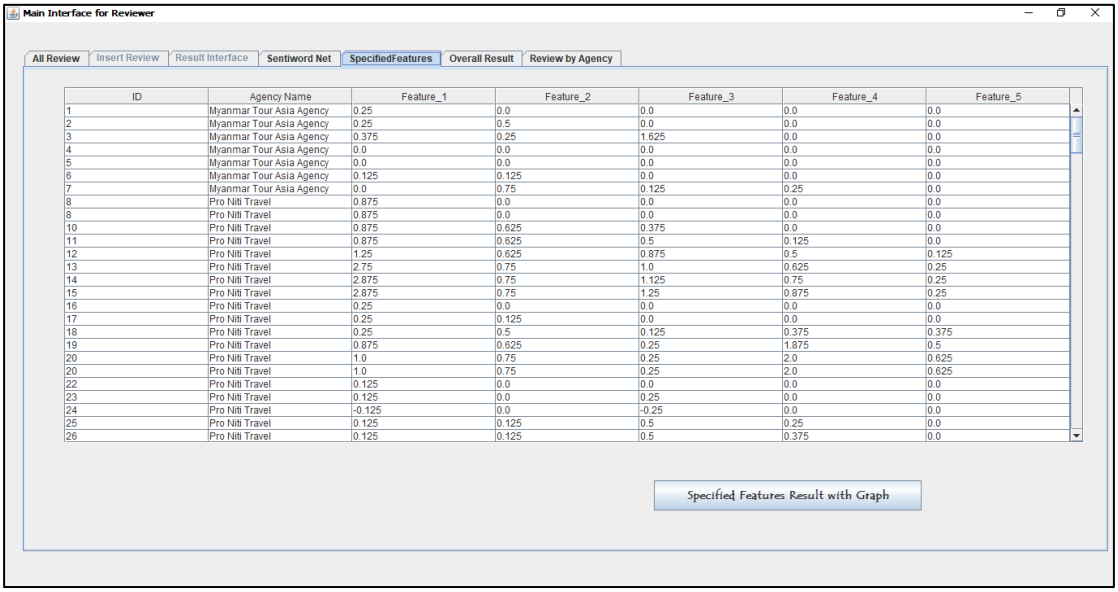
The system automatically calculates the final weight of each trip review and shows the total weight of positive features and the total weight of negative features.



POS	ID	Positive_Score	Negative_Score	Synset_Terms	Gloss
a	1740	0.125	0.0	able#1	(usually followed by `to') having th...
a	2098	0.0	0.75	unable#1	(usually followed by `to') not havin...
a	2312	0.0	0.0	dorsal#2 abaxial#1	facing away from the axis of an or...
a	2527	0.0	0.0	ventral#2 adaxial#1	nearest to or facing toward the axi...
a	2730	0.0	0.0	acroscopic#1	facing or on the side toward the a...
a	2843	0.0	0.0	basiscopic#1	facing or on the side toward the b...
a	2956	0.0	0.0	abducting#1 abducent#1	especially of muscles; drawing a...
a	3131	0.0	0.0	adductive#1 adducting#1 adduce...	especially of muscles; bringing to...
a	3356	0.0	0.0	nascent#1	being born or beginning; the nasc...
a	3553	0.0	0.0	emerging#2 emergent#2	coming into existence; an emerg...
a	3700	0.25	0.0	dissilient#1	bursting open with force, as do s...
a	3829	0.25	0.0	parturient#2	giving birth; a parturient heifer
a	3939	0.0	0.0	dying#1	in or associated with the process ...
a	4171	0.0	0.0	moribund#2	being on the point of death; breat...
a	4296	0.0	0.0	last#5	occurring at the time of death; his ...
a	4413	0.0	0.0	abridged#1	(used of texts) shortened by cond...
a	4615	0.0	0.0	shortened#4 cut#3	with parts removed; the drasticall...
a	4723	0.0	0.0	half-length#2	abridged to half its original length
a	4817	0.0	0.0	potted#3	(British informal) summarized or ...
a	4980	0.0	0.0	unabridged#1	(used of texts) not shortened; an ...
a	5107	0.5	0.0	uncut#7 full-length#2	complete; the full-length play
a	5205	0.5	0.0	absolute#1	perfect or complete or pure; absol...
a	5473	0.75	0.0	direct#10	lacking compromising or mitigati...
a	5599	0.5	0.5	unquestioning#2 implicit#2	being without doubt or reserve; i...
a	5718	0.125	0.0	infinite#4	total and all-embracing; Gods infi...
a	5839	0.5	0.125	living#3	(informal) absolute; she is a livin...

Figure 4. 15 SentiWordNet 3.0

The positive score and negative score of each opinion word have been shown in SentiWordNet 3.0. There are many opinion words in this SentiWordNet 3.0 so that reviewer can get correct information.



ID	Agency Name	Feature_1	Feature_2	Feature_3	Feature_4	Feature_5
1	Myanmar Tour Asia Agency	0.25	0.0	0.0	0.0	0.0
2	Myanmar Tour Asia Agency	0.25	0.5	0.0	0.0	0.0
3	Myanmar Tour Asia Agency	0.375	0.25	1.625	0.0	0.0
4	Myanmar Tour Asia Agency	0.0	0.0	0.0	0.0	0.0
5	Myanmar Tour Asia Agency	0.0	0.0	0.0	0.0	0.0
6	Myanmar Tour Asia Agency	0.125	0.125	0.0	0.0	0.0
7	Myanmar Tour Asia Agency	0.0	0.75	0.125	0.25	0.0
8	Pro Niti Travel	0.875	0.0	0.0	0.0	0.0
8	Pro Niti Travel	0.875	0.0	0.0	0.0	0.0
10	Pro Niti Travel	0.875	0.625	0.375	0.0	0.0
11	Pro Niti Travel	0.875	0.625	0.5	0.125	0.0
12	Pro Niti Travel	1.25	0.625	0.875	0.5	0.125
13	Pro Niti Travel	2.75	0.75	1.0	0.625	0.25
14	Pro Niti Travel	2.875	0.75	1.125	0.75	0.25
15	Pro Niti Travel	2.875	0.75	1.25	0.875	0.25
16	Pro Niti Travel	0.25	0.0	0.0	0.0	0.0
17	Pro Niti Travel	0.25	0.125	0.0	0.0	0.0
18	Pro Niti Travel	0.25	0.5	0.125	0.375	0.375
19	Pro Niti Travel	0.875	0.625	0.25	1.875	0.5
20	Pro Niti Travel	1.0	0.75	0.25	2.0	0.625
20	Pro Niti Travel	1.0	0.75	0.25	2.0	0.625
22	Pro Niti Travel	0.125	0.0	0.0	0.0	0.0
23	Pro Niti Travel	0.125	0.0	0.25	0.0	0.0
24	Pro Niti Travel	-0.125	0.0	-0.25	0.0	0.0
25	Pro Niti Travel	0.125	0.125	0.5	0.25	0.0
26	Pro Niti Travel	0.125	0.125	0.5	0.375	0.0

Figure 4. 16 Specified Features of Each Review

To show the specified features for each review, the system automatically generate the five specified features for each review using term frequency-inverse document frequency. In each review, there are many features which are mentioned by the reviewer. Therefore, the system produces only specified features using IF_IDF.

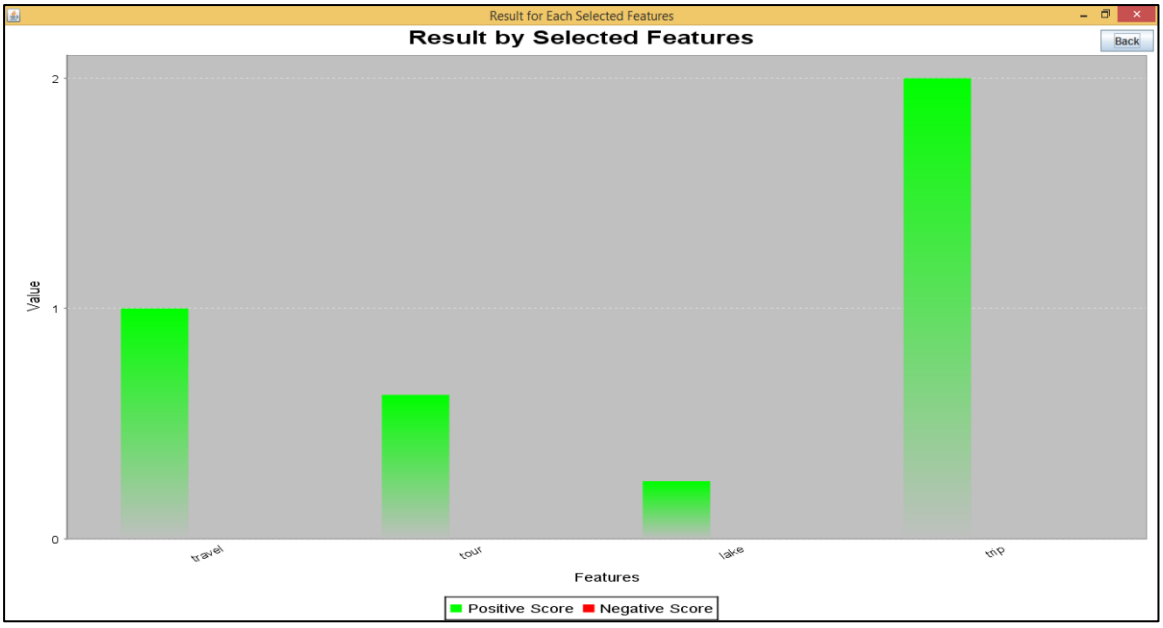


Figure 4. 17 Result by Selected Features Graph

In this graph, the system shows the positive features with the green color and the negative features with the red color.

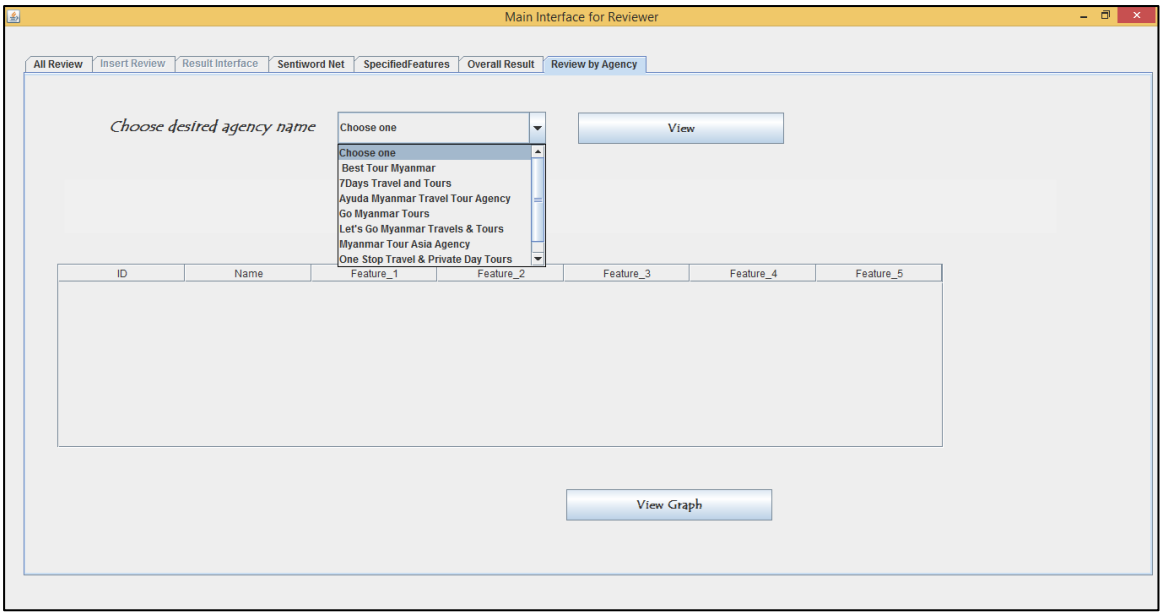


Figure 4. 18 Final Result of Five Specified Features by Each Agency

If the reviewer wants to search for each agency review, the reviewer can choose an agency that the reviewer wants to search and click the view button.

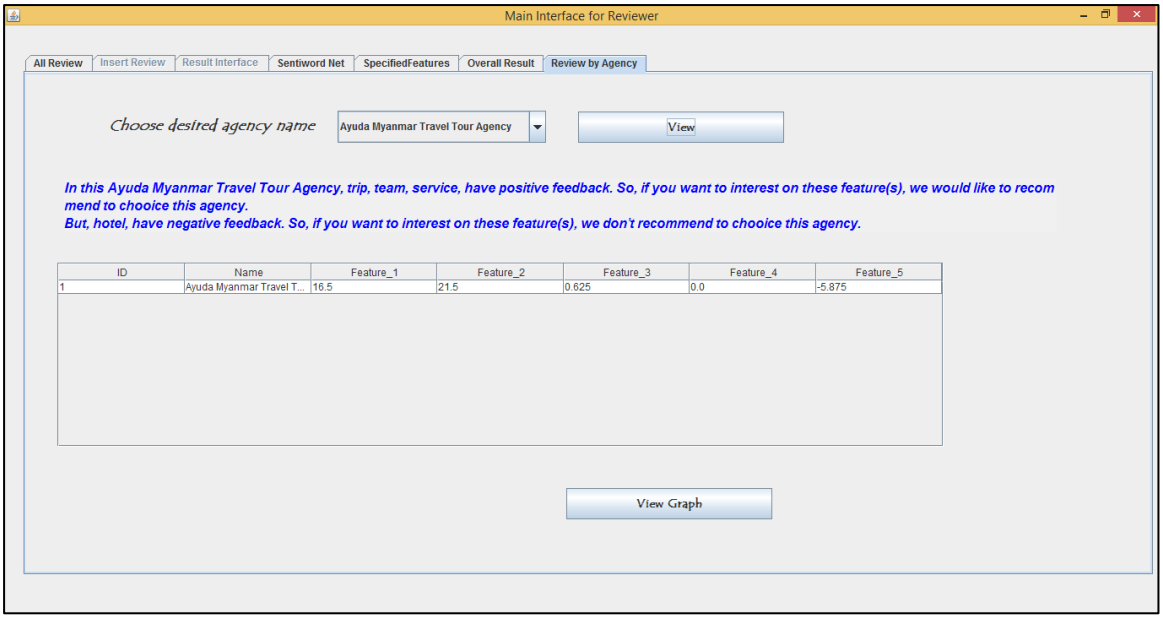


Figure 4. 19 Five specified Features Values by Each Agency

Then, the system will show the five specified features that are frequently mentioned by the reviewer. If the reviewer wants to view with graph, the reviewer must click View Graph button.

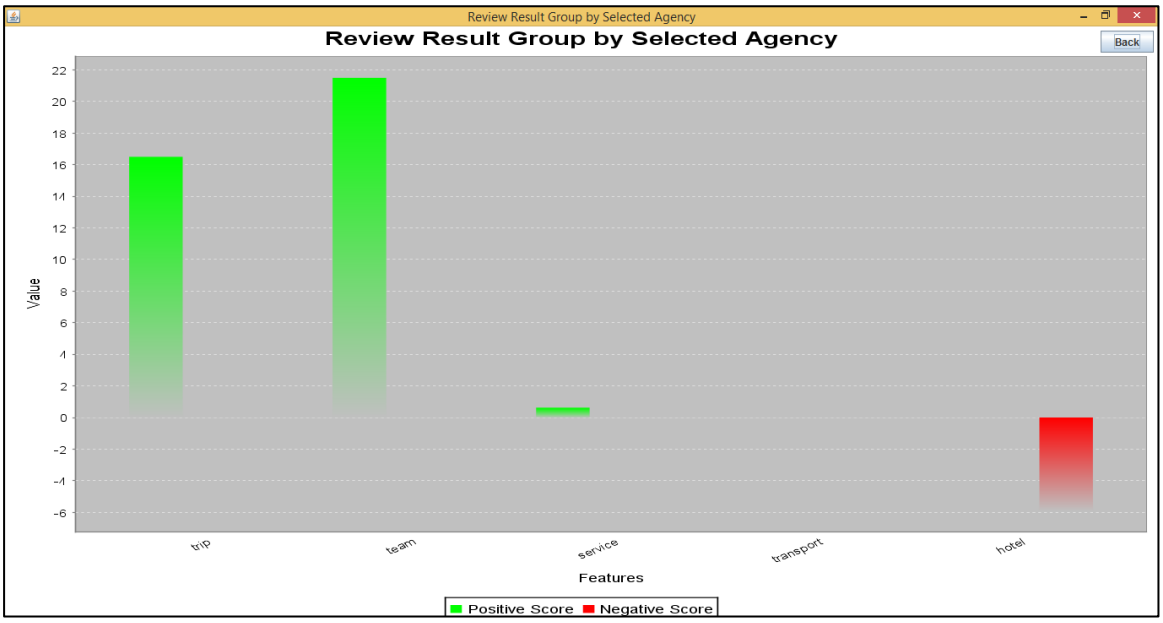


Figure 4. 20 Review Result Group by Selected Agency

This graph shows the most frequently specified features by the reviewer with the green color for the positive value and the red color with negative value.

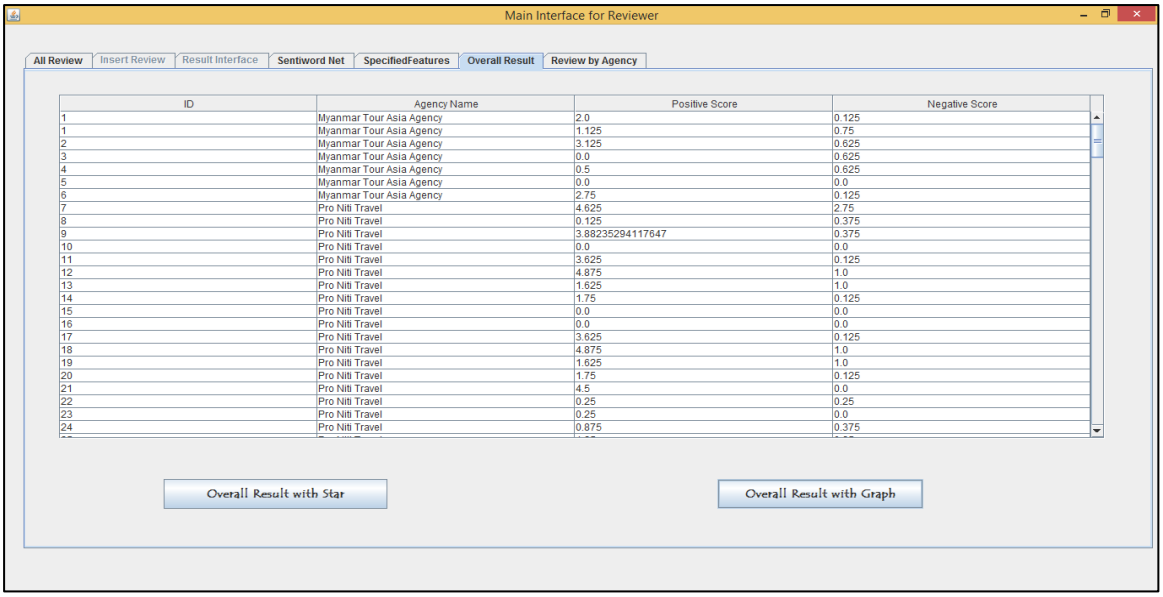


Figure 4. 21 The Total Weight of Posscore and Negscore of Each Review

The system shows the total weight of positive score and negative score for each review. If the reviewer clicks the overall result with graph button, the system will show the figure as following.

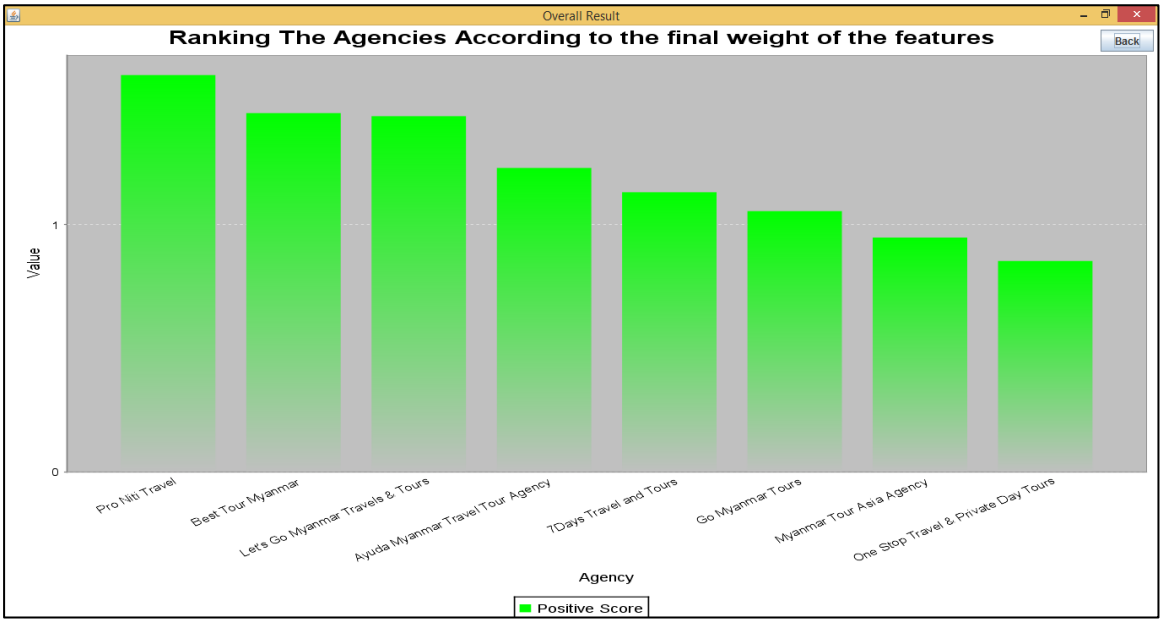


Figure 4. 22 Ranking the Agencies According to the Final Weight of the Features

Finally, the system shows the overall final result of each agency and the system automatically shows each agency with the ranking of final score. The above figure shows the raking all agencies according to the total weight of the features.

By clicking the Viewer button, the system will show the Main Interface for Viewer.

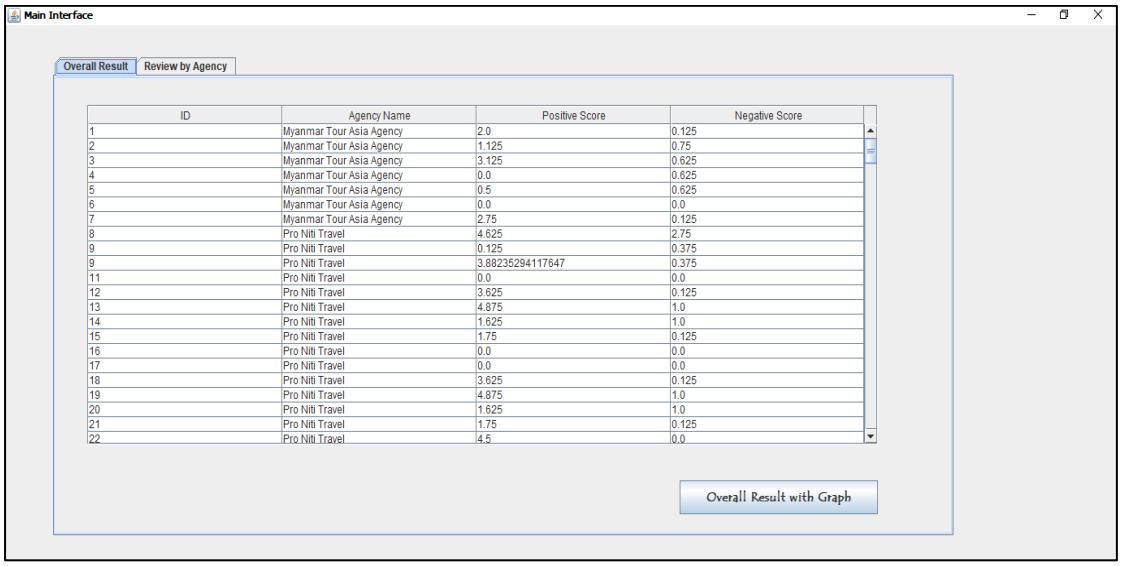


Figure 4. 23 Main Interface for Viewer

This is the main interface for the viewer. In this viewer main interface, there are two menus such as Overall Result and Review by Agency because viewer does not need to insert the review and he/she needs to view the overall result and review by agency.

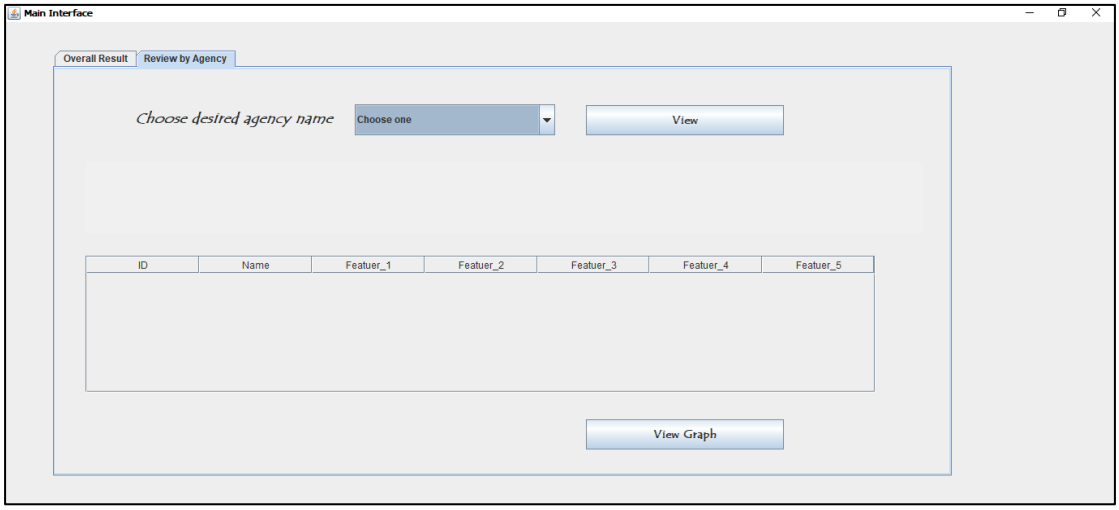


Figure 4. 24 Final Result of Five Specified Features by Each Agency

If the viewer wants to search for each agency review, the viewer can choose an agency that the viewer wants to search and click the view graph button.

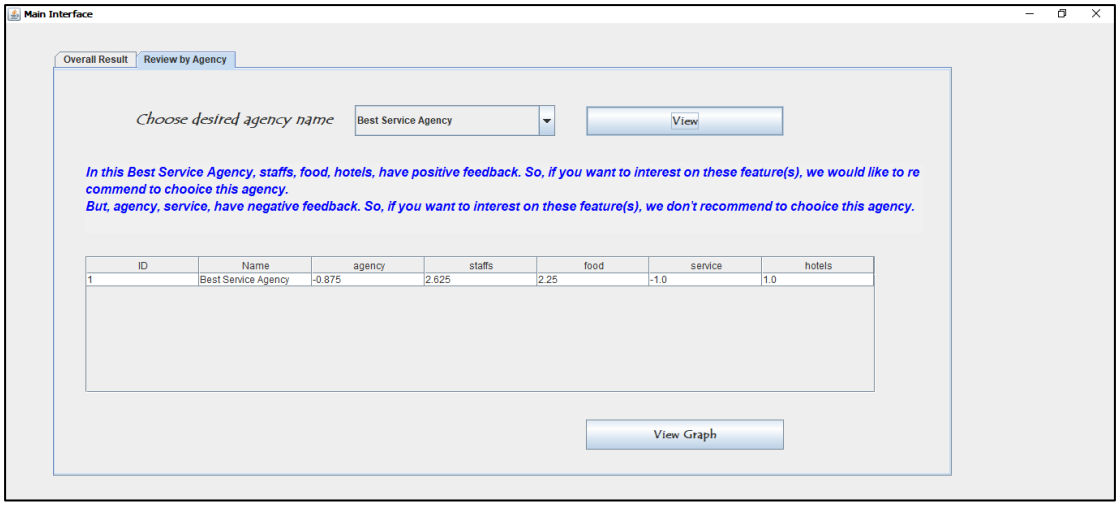


Figure 4. 25 Five Specified Features Values by Each Agency

Then, the system will show the five specified features that are frequently mentioned by the reviewer. If the viewer wants to view with graph, the viewer must click View Graph button.

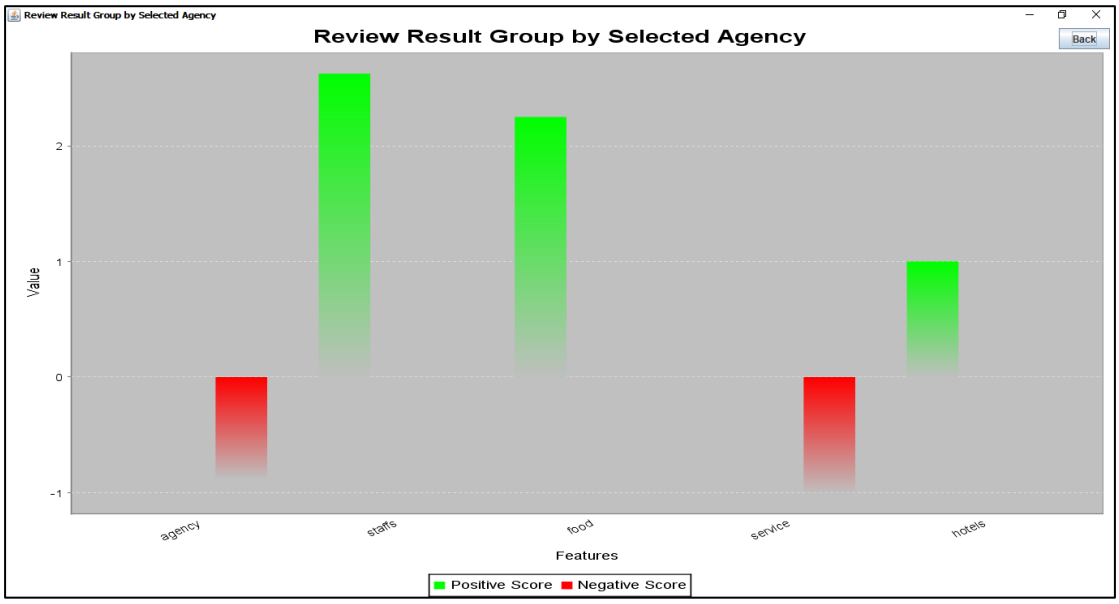


Figure 4. 26 Review Result Group by Selected Agency

This graph shows the most frequently specified features by the viewer with the green color for the positive value and the red color with negative value.

The system have been tested using the complex positive and negative sentences instead of the simple positive and negative sentences such as “don’t mean”, “never”, “don’t/doesn’t like”, “not..not”, “never mind” and so on.

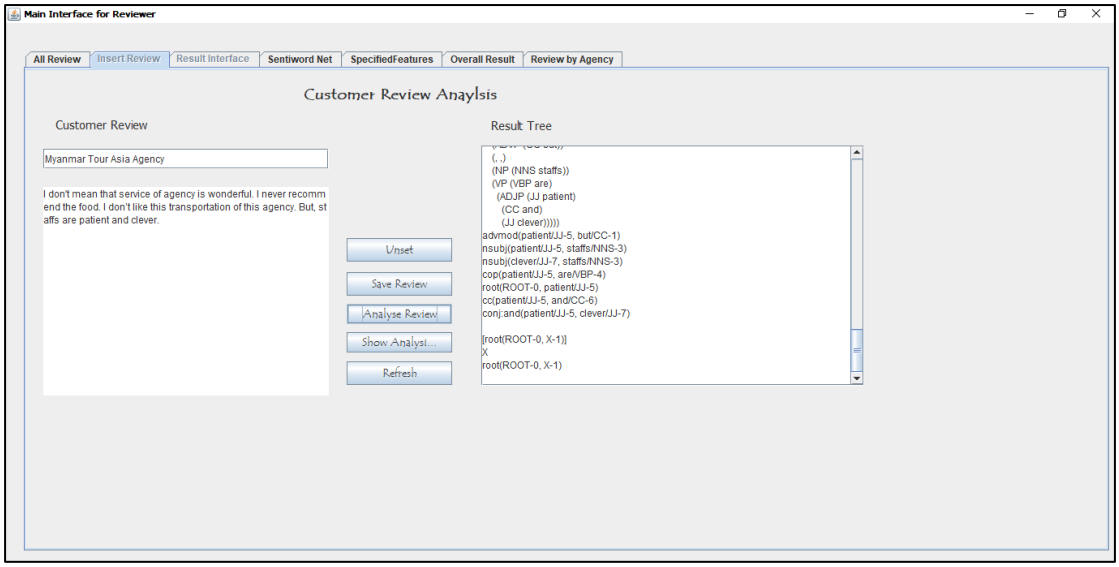


Figure 4. 27 Testing the Complex Positive and Negative Sentences

The system used the “I don’t mean that the service of the agency is wonderful” instead of “The service of the agency is bad/ not wonderful/ unsatisfied”. “The service of the agency is wonderful” is positive review but “I don’t mean that the service of the agency is wonderful” is negative review because of “don’t mean”. The other sentences are the same with the above sentence.

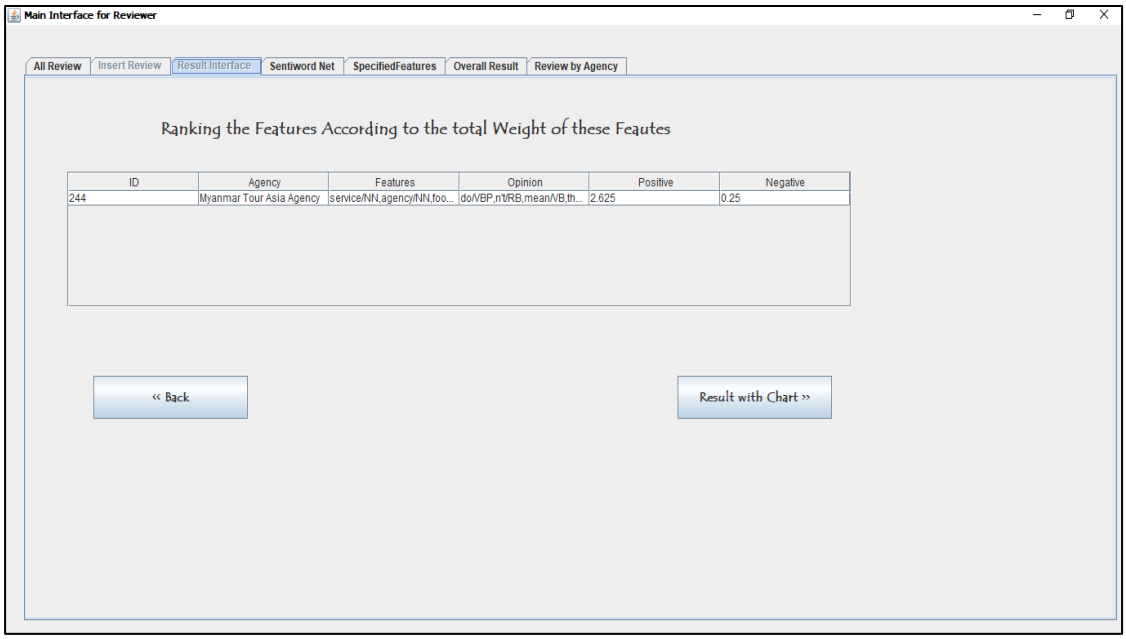


Figure 4. 28 Final Weight of a Trip Review

The system automatically calculated the final weight of each trip review and shows the total weight of positive features and the total weight of negative features.

Figure 4. 29 Result by Selected Features

After clicking the "Result with Chart>>" button, the system shows the result by selected features with posscore and negscore.

4.7 Performance Analysis

The performance analysis of the proposed system can be measured by computing its efficiency and its effectiveness. This system measures accuracy result using F1 score. F1 score (also F1 score or F-measure) is a measure of a test's accuracy. It considers both the Precision and the Recall of the test to compute the score. Precision and recall are defined in terms of a set of retrieved documents (eg. the lists of documents produced by a search engine for a query) and a set of relevant documents (eg. the lists of all documents that are relevant for a certain topic). The equation for calculating F1 score can be seen in the following equations (4.1), (4.2), (4.3) and (4.4).

$$F1=2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precison+Recall}} \qquad 4.1$$

$$\text{Accuracy} = \frac{\text{True Positive+True Negative}}{\text{True Positive +False Negative+True Negative+ False Positive}} \qquad 4.2$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive + False Positive}} \quad 4.3$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive + False Negative}} \quad 4.4$$

This system is tested by using trip reviews. First, the reviews are analyzed by using sentiment analysis module. Then, the output results are kept and checked with the manually pre-classified positive and negative reviews. After testing positive pre-classified positive and negative reviews, the accuracy result can be seen in Table 4.4.

Table 4.4 Evaluation Result of the System

		Actual	
		Positive Reviews	Negative Reviews
Prediction	Positive Reviews	170 (TP)	13 (FP)
	Negative Reviews	4 (FN)	25 (TN)

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{170}{170+13} = \frac{170}{183} = 0.93$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{170}{170+4} = \frac{170}{174} = 0.98$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precison+Recall}} = 2 \times \frac{0.93 \times 0.98}{0.93+0.98} = 0.95$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+TN+FP} = \frac{170+25}{170+4+25+13} = \frac{195}{212} = 0.92$$

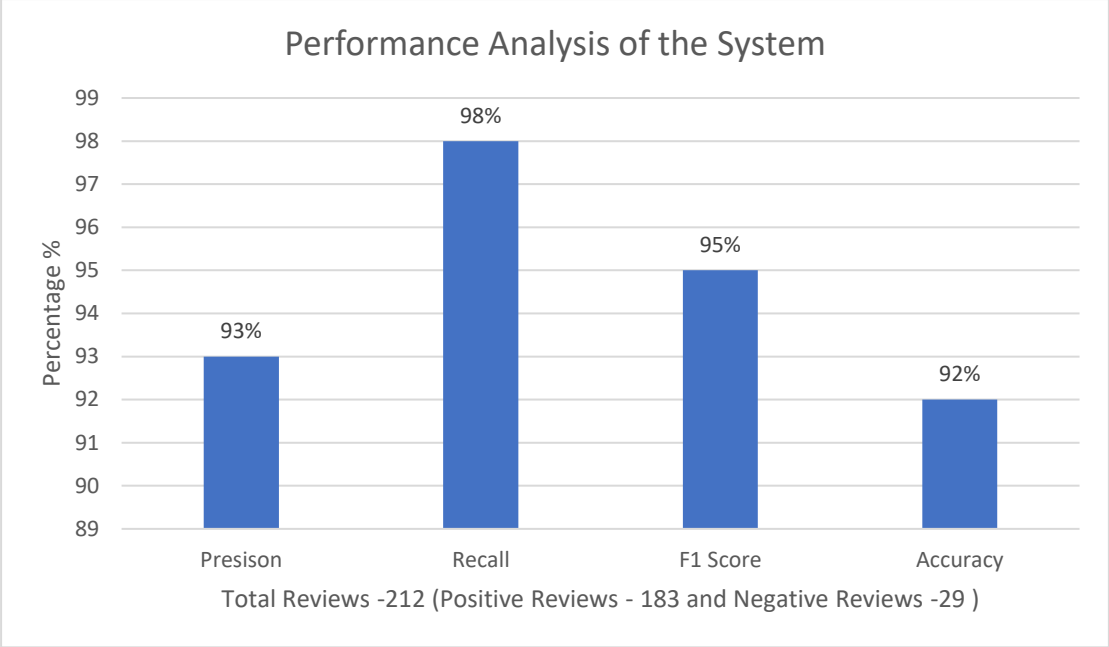


Figure 4. 30 Performance Analysis of the System

4.8 Comparing the System Result with the Tripadvisor

After finishing the whole system, the system is needed to compare with the real Tripadvisor website because the system used the trip review sentences from the Tripadvisor. There is a little difference between the following two graphs. The result from Tripadvisor is calculated by using the latest or up to date trip reviews that are posted by the reviewers who are recently travelled with these travel agencies. Although there are many trip reviews on the Tripadvisor, there are some trip reviews on the system. If the reviewers post the up to date trip reviews on the system with the same time, the system will calculate the correct and up to date information about these agencies. On the Tripadvisor website, if the reviewers want to know about the agencies, they have to read many trip reviews and then they cannot exactly know which features are reviewers like or dislike. But, the system can calculate the exactly information about these features within a minute.

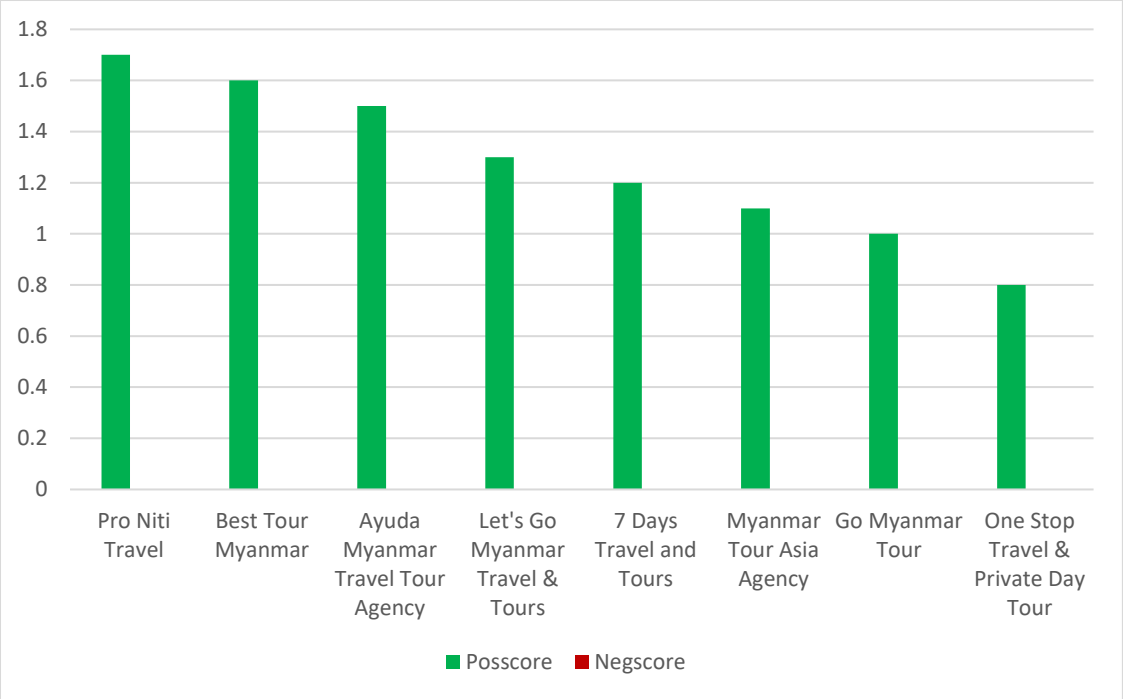


Figure 4. 31 Result from the Tripadvisor Website

Figure 4.30 shows the final result from the Tripadvisor website that are written by the reviewers.

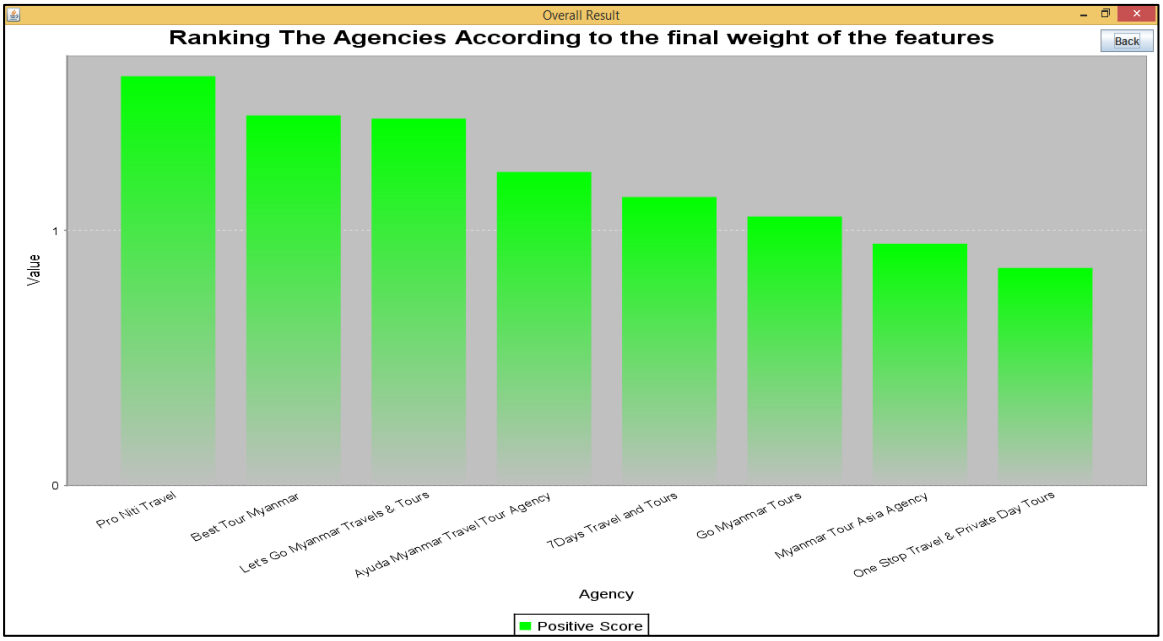


Figure 4. 32 Result from the system

Figure 4.31 shows the final result that are calculated by the system.

CHAPTER 5

CONCLUSION

This system intended to develop a feature extraction system from tourism domain. This system will be parsed the review sentences and identified the features efficiently, and then the weight of frequent features will be obtained and ranked these features according to their score values. As the system showed the result as ranking, customers and administrators would know the trip features which are generally liked and disliked by the customer. Therefore, customer can get valuable facts about tour agent they want. Moreover, each tour agency can directly know the strength and weakness of theirs so that trip agency can target in those trip features.

5.1 Advantages of the System

This system has implemented opinion mining using SentiWordNet 3.0. This system not only can be used in any kind of trip reviews but also agency reviews and then can be analyzed the reviews. The analyzed reviews are useful to high level trip managers. Therefore, this system is used for the important decision-making process. This system can be used trip reviews as well as other reviews by modifying features.

5.2 Further Extension

In future, some extensions are proposed to increase the capabilities and efficiency of Data Mining System. The system has been tested on tourism reviews form Tripadvisor website for new customers or travellers. It can be extended to use any dataset for marketing such as cosmetic reviews for beauty bloggers, movie recommender system, and so on by using feature extraction and SentiWordNet3.0.

5.3 Limitations of the System

Most of the reviewers have not followed the grammatical rules when writing reviews in the proposed system can miss some opinion words. As a result, the errors come from the syntactic parser and dependency link and implicit opinion expressions

and typo cannot be got good the precision value. Therefore, reviewers should follow the grammatical rule in this system.

Moreover, the system shows the five specified features that are frequently mentioned by the reviewers in the reviews because the system cannot show all features from the reviews. These five specified features are not up to date because the system automatically calculated the top five specified features based on the existing reviews from the database using IF_IDF formula. If reviewers want to get the correct and up to date review information, reviewers should insert the latest and up to date reviews from the Tripadvisor website.

REFERENCES

[1] Abbasi, Ahmed, Hsinchun Chen, and Arab Salem. "Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums". ACM ransactions on Information Systems (TOIS), 2008.

[2] Abdul-Mageed, Muhammad, Mona T. Diab, and Mohammed Korayem. "Subjectivity and sentiment analysis of modern standard Arabic". In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers. 2011.

[3] Alec, G.; Lei, H.; and Richa, B. Twitter sentiment classification using distant supervision. Technical report, Standford University. 2009

[4] Alekh Agarwal and Pushpak Bhattacharyya, Sentiment Analysis: A New Approach for Effective Use of Linguistic Knowledge and Exploiting Similarities in a Set of Documents to be Classified, International Conference on Natural Language Processing (ICON 05), IIT Kanpur, India, December, 2005

[5] Alok Choudhary, Zhang K., Narayanan R., and Choudhary A., "Mining Online Customer Reviews for Ranking Products", Technical Report, EECS department, Northwestern University, (2009)

[6] Ana-Maria Popesu and Oren Etzioni "Extracting Product Features and Opinions from Reviews", Proceeding of Human Language Technology Conference and Conference on Empirical Methods in Natural Lagunage, ACL, Vancouver, pp. 339-346, , (2005)

[7] Andrea Esuli and Fabrizio Sebastiani, (2006), "SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining", In Proceedings of LREC-

06, 5th Conference on Language Resources and Evaluation, Genova, IT, pp. 417-422.

[8] A.-M. Popescu and O. Etzioni, “Extracting product features and opinions from reviews,” in Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), 2005.

[9] B. Liu, “Opinion mining and sentiment analysis,” Handbook of Natural Language Processing, 2010.

[10] B. Pang and L. Lee, “Opinion mining and sentiment analysis,” Foundations and Trends in Information Retrieval, vol. 2, no. 1-2, pp. 1–135, 2008.

[11] Cristian Bucurab “Using Opinion Mining Techniques In Tourism”,2nd Global Conference On Business, Economics, Management And Tourism, 30-31 Oct, 2014, Prague, Czech Republic.

[12] David N. Milne and Ian H. Witten and David M. Nichols, “A knowledge-based search engine powered by wikipedia, Proceedings of the Sixteenth ACM conference on Conference on information and knowledge management”, ACM New York, NY, USA, 2007

[13] Deepanshi Sharma, Achal Kulshreshtha, Priyanka Paygude,” Tourview: Sentiment Based Analysis On Tourist Domain”, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (3), 2015, 2318-2320.

[14] Dey, Lipika and Haque, Sk. “Opinion Mining from Noisy Text Data, International Journal on Document Analysis and Recognition” 12(3). pp 205-226, 2009

[15] Edmundson, H. P., New methods in automatic extracting, Journal of the ACM, 16(2): 264-285, 1969

[16] Ekenel, Hazim Kemal and Semela, Tomas and Stiefelhagen, Rainer, "Content-based video genre classification using multiple cues, Proceedings of the 3rd international workshop on Automated information extraction in media production", AIEMPro '10, 2010

[17] Elwell, Robert and Baldridge, Jason, "Discourse Connective Argument Identification with Connective Specific Rankers", In Proceedings of IEEE International Conference on Semantic Computing, 2008

[18] Esuli A, Sebastiani F., SentiWordNet: "A Publicly Available Lexical Resource for Opinion Mining", In Proceedings from International Conference on Language Resources and Evaluation (LREC), Genoa, 2006

[19] Giuliano Armano and Alessandro Giuliani and Eloisa Vargiu, "Experimenting Text Summarization Techniques for Contextual Advertising", Proceedings of the 2nd Italian Information Retrieval (IIR) Workshop, Milan, Italy, 2011

[20] Grishman, R., "Adaptive information extraction and sublanguage analysis", In Proceedings of the 17th International Joint Conference on Artificial Intelligence, 2001

[21] Himabindu Lakkaraju, Chiranjib Bhattacharyya, Indrajit Bhattacharya and Srujana Merugu, Exploiting Coherence for the simultaneous discovery of latent facets and associated sentiments, SIAM International Conference on Data Mining (SDM), April 2011 44. Hirst, G. & St-Onge, D., 'Lexical chains as representation of context for the detection and correction malapropisms, 1997

[22] Kunpeng Zhang Ramanathan Narayanan, “Voice of the Customers: Mining Online Customer Reviews for Product Feature-based Ranking”, Electrical Engineering and Computer Science Department Northwestern University, (2010)

[23] N. Hu, P. Pavlou, and J. Zhang, Can Online Reviews Reveal a Product’s True Quality? Empirical Findings and Analytical Modeling of Online Word-of-Mouth Communication, EC., 6 (2006), pp. 324-330.

[24] Qi Zhang, Yuanbin Wu, Tao Li, Mitsunori Ogihara, Joseph Johnson, Xuanjing Huang,”Mining Product Reviews Based on Shallow Dependency Parsing”, SIGIR '09, Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, 2009.

[25] S. Aciar, “Mining context information from consumer’s reviews,” in Proceedings of the Context-Aware Recommender Systems (CARS) Workshop, 2009.

[26] T. Wilson, J. Wiebe, and P. Hoffmann, “Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis,” Computational Linguistics, 2005.

[27] Wang, Pu and Domeniconi, Carlotta, “Building semantic kernels for text classification using Wikipedia”, Proc. of SIGKDD, 2008

[28] Webber, Bonnie and Knott, Alistair and Stone, Matthew and Joshi, Aravind, “Discourse relations: A structural and presuppositional account using lexicalized tag”, In Proceedings of ACL, 1999

[29] Wellner, Ben and Pustejovsky, James and Havasi, Catherine and Rumshisky, Anna and Saur , Roser, “Classification of discourse coherence relations: an

exploratory study using multiple knowledge sources", Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue, 2006

[30] Weishu Hu, Zhiguo Gong, Jingzhi Guo, "Mining Product Features from Online Reviews", Faculty of Science and Technology University of Macau , China, (2010)

[31] Wolf, F., Gibson, E. and Desmet, T., Discourse coherence and pronoun resolution, Language and Cognitive Processes, 19(6), pp. 665–675, 2004

[32] Wolf, Florian and Gibson, Edward, Representing discourse coherence: A corpus-based study, Computational Linguistics, 31(2), pp. 249–287, 2005