



S	
E1	
E2	
For Office Use Only	

Masters Project Final Report
(MCS)
2019

Project Title	Aspect based sentimental analysis of Sri Lankan Hotels
Student Name	M. Gunawardana
Registration No. & Index No.	2014/MCS/026 - 14440263
Supervisor's Name	Dr. H.A. Caldera

For Office Use ONLY



Aspect based sentimental analysis of Sri Lankan Hotels

**A dissertation submitted for the Degree of Master of
Computer Science**

M.Gunawardana

University of Colombo School of Computing

2019



Declaration

The thesis is my original work and has not been submitted previously for a degree at this or any other university/institute.

To the best of my knowledge it does not contain any material published or written by another person, except as acknowledged in the text.

Student Name: M. Gunawardana

Registration Number:2014/MCS/026

Index Number:14440263

Signature:

Date:

This is to certify that this thesis is based on the work of Ms. Mihirani Gunawardana under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by:

Supervisor Name: Dr. H.A. Caldera

Signature:

Date:

Abstract

Now a day's people use to check review comments before they are going to experience new things most of the time. This is quite common specially in hotel areas. These reviews are the most convenient way for tourists to get an idea about the place they are going to visit and stay. Not only tourists, local people and service providers also refer these reviews. But the problem is people are so busy with day to day life and they don't have much time to check the review manually. Specially people like hotel managers. So, people use to read top most reviews and get the idea. Most of the time that is not effective. Sometimes we get the idea that the place is good when we go through the first few reviews, but if we consider the overall comments it can be different. So, there should be an easy way to find this overall idea. For hotel management also can do their job easily if there is an automatic way to find about the status about their hotels.

This research will try to provide a solution for the above-mentioned problem. Most of the text review provide by the customer are unstructured and not organized in a pre-defined manner. These texts are usually difficult, time-consuming and expensive to analyze, understand, and sort through. This is a complex analysis. These kinds of complex analysis can be done by using Sentiment analysis. It is also known as opinion mining. In sentiment analysis, unstructured information could be automatically transformed into structured data of public opinions. This is the most suitable one for mine the opinion of human. It is a type of text analysis that classifies the human generated text and makes decision by extracting and analyzing the text. It extracts the hidden knowledge from the unstructured texts exist in form of patterns and relationships. Opinion can be categorized as positive and negative, then measure the degree of positive or negative associated with the event.

There are three type of opinion mining techniques. Document level, Sentence level and Feature /Aspect level. In document level it considers whole document or paragraph and provide the entire document or paragraph looks more positive than negative. In sentence level, it breaks the document into sentences and classify each of them as positive, negative or neutral. This is bit more specific here than document level. In Aspect level it is more specific than other two level. It extracts the specific feature and check the positivity or negativity according to that.

If we consider the review comments. User put those comments considering some specific features. As an example, if we consider hotel field, customer provide review considering food or rooms or other aspect. Customer or service provider who read these reviews also read considering these kinds of aspect. So, they need specific idea. If they want to know about food of that hotel. It is no point to provide the entire document or sentence level idea. Sometime entire sentence or document can be positive but idea about food can be negative. So, method use to analyze customer review should more specific. Because of that, in this research it used aspect level opinion mining.

Acknowledgement

Firstly I would like to express my sincere gratitude to my mentor Dr.H.A.Caldera for his immense support in completing my research. His continuous guidance and knowledge helped me a lot to improve the quality of my research.

I should mention the library of UCSC and the staff for helping me to find out related research papers.

Last but not the least I would like to thank my family and friends for their unfailing support and continuous encouragement through these years. This accomplishment would not have been possible without them.

Contents

Introduction	10
1.1. Problem.....	12
1.2. Motivation.....	13
1.3 The exact computer science problem.....	14
1.4 Objective	14
1.5 Scope.....	14
1.6 Summary	14
Literature Review and Background.....	15
2.1. Introduction	15
2.2. Related Works.....	16
Using Opinion Mining Techniques in Tourism	16
Mining Opinion Features in Customer Reviews.....	16
Aspect based Opinion Mining from Restaurant Reviews	17
Aspect Based Sentiment Analysis to Extract Meticulous Opinion Value	18
Sentiment analysis of movie reviews.....	19
2.3 Sentimental analysis (Opinion Mining)	19
2.4 Summary	20
Design and Methodology.....	21
3.1. Introduction	21
3.2. Research methodology	21
3.3. System Architecture.....	21
3.3.1. Data Collection.....	23
3.3.2. Preprocessor	23
3.3.3. Tokenize and Aspect extraction (POS Tagging).....	24
3.3.4. Recognize Opinion word (Dependency Parser)	26
3.3.5. Get Sentiment Score (SentiwordNet).....	28
3.4. Problem Analysis.....	30
3.4 Summary	31
Implementation	32
4.1. Introduction	32
4.2. Preprocessing.....	32
4.3. Tokenizing and Aspect extraction (POS tagging)	33
4.4. Dependency Parsing.....	37
4.5. Sentiment Score Calculator (SentiWordNet)	39
Results and Evaluation	41

5.1. Introduction	41
5.2. Manual Test phase	41
5.2.1 First manual phase	41
5.2.2 Second manual phase	43
5.3 Final Evaluation	45
5.4. Summary	47
Conclusion and Future works	48
6.1. Conclusion	48
6.2 Future Work	49
References	50

List of Abbreviations

Abbreviations	Explanation
POS	Part-Of-Speech
NN	Noun
NNS	Noun plural
NNP	proper noun, singular
Synset	English words into sets of synonyms

Chapter 1

Introduction

In past time people use traditional channel to know details about something they want to know. Most of the time tourists got information mainly through newspapers, broad cast and paper materials from travel agencies. And the information they got also limited to some area like tourism route, travel vehicles and prices. But that is not enough for them.

Nowadays using of internet and social media is growing very rapidly. People also use these social media and internet to share their experiences and ideas. They use different methods to share their experiences. Some put their experiences and ideas as a review comment, and some put their idea as a post in social media. Most people use to read these review comments before deciding something. Based on this there are many recommendation systems appeared on the society. Food recommendation system, travel recommendation system, Health tips recommendation system, product recommendation system and there are so many. These all systems use the details that people published through the internet, for generating results.

Among all those, travel recommendation gets high priority when it comes to the tourism. Most of the tourist don't forget to leave their comments before they left. Those time people maintain a log book to collect comments. But now there is site for every hotel and sometimes some other third party maintained a site on behalf of these hotels. Trip advisor is a such kind of site. People can leave their comment on that site.

When go through that site there is lot of comments. To find these comments also get time. Going through all the review about one hotel, it takes hours. According to the data collected, one hotel has 1716 review comments. Some comments have more than 30 lines. So, reading this manually is not easy. People are also busy with day to day life and they don't have much time to check the review manually. Specially people like hotel managers. They don't have time to go through each comment and analysis them. So, if there is automatic way to read and analysis, it can be very useful to both service provider and the customer.

Most of the text review provide by the customer are unstructured and not organized in a pre-defined manner. These texts are usually difficult, time-consuming and expensive to analyze, understand, and sort through. This is a complex analysis. These kinds of complex analysis

can be done by using Sentiment analysis. It is also known as opinion mining. In sentiment analysis, unstructured information could be automatically transformed into structured data of public opinions. This is the most suitable one for mine the opinion of human. It is a type of text analysis that classifies the human generated text and makes decision by extracting and analyzing the text. It extracts the hidden knowledge from the unstructured texts exist in form of patterns and relationships. Opinion can be categorized as positive and negative, then measure the degree of positive or negative associated with the event.

Sentimental analysis can be done at tree level.

- Document level
- Sentence level
- Feature /Aspect level

These three are discuss later in this document with more details. For this research Aspect level is the most suitable method. Let take an example from TripAdvisor and discuss about this.

Following is one review that is mentioned in the TripAdvisor site about a hotel.

“The views from our balcony, restaurant and lounge deck are the best to be had in all of Ella! It is a pleasant 5 minutewalk to the centre of town and a one-minute walk to laundry facilities and cooking class. The owner is knowledgeable, helpful and courteous as is all the staff. Breakfast was plentiful and delicious. We highly recommend a stay at Green Hill.

Room Tip: Thee double room with balcony is a great room to book.”

In here customer mentioned their opinion based on different aspects like view, laundry facility, staff, food. And she also gives a room tip. These kinds of data must be analyzed on feature level. So, document level and sentence level are not the techniques suitable for hotel review analysis. At both levels it does not discover what people like or not.

1.1. Problem

Growth of using internet encourage people to use it frequently. So, people share their experiences day by day through internet. They use internet to publish their experience as review comments. Most of the travelers share their experience about their journey in their blogs or in travel details web sites based on many aspects. Other travelers read these reviews before they get a decision about a place. Because of that, review comments pool is also growing rapidly. When there is lot of data in the review pool, it is not an easy task to go through each comment. It is a time-consuming process. So, reading and analyzing this kind of large data manually is not easy. Sometimes it is impossible. Because of that, they need some easy way to find out all details.

Travelers aspects can be different. Some expects good foods or amazing sceneries, some expects comfortable rooms for reasonable prices ... etc. So going through lot of review comments and find these aspects are not an easy task. Sometimes it takes times to find review comment also. Sometimes in first few comments there won't be the things that travelers looking for. Because of that they tend to look another even though the suitable one is the previous one.

From management point of view. They are very busy with the day to day activity and they don't have time to look at reviews comments and going through each comment. They also want to know the status based on some aspects. As an example, if they want to improve the quality of food, they are looking comment towards foods. So, going through all comments is pointless. If there is no easy way, then management have to maintain separate staff to maintain review as well.

Already there is some rating methodologies available in these travel details websites, but it is not enough to get a good decision. When consider the above problems there should be a easy effective way. If there is way to see all these as summary as they expected, it will be very useful to them. If this information can be summarized and categorized using aspect based sentimental analysis, it will be very important and useful.

1.2. Motivation

As a habit, now people refer other's opinion about things before they get final decision. Before buying a product, they will go through review comments. Growing usage of internet encourages people to share their opinion about these product and services publicly and that helps others to get idea about them. This helps customers to find a good product and services and also helps service providers. Service providers can identify the competitors and what they provide better than them. That will encourage them to improve their product or service.

Hotel field is similar to this. After visiting places people put their comments on internet. So, others can refer them and can get an idea about these places before getting a decision to visit there or not. But review pool is getting larger and larger since people post their comments every day. So, going through each comment manually is not an easy task. It is very time consuming. When we consider hotel field people's expectations are on different aspects. So, they have to go through every review comment to get ideas. It is better if they can see these as a summary according to their aspects. Then they can decide very quickly.

Not only the hotel customers, Hotel management also refer these reviews to improve the quality of their service. So, they also get ease of doing things if there is a summary system like above mention.

As an example, suppose some visitor wants to find a hotel in Sri Lanka which has good food, beautiful views and excellent service for a reasonable price. Now they want to find a review which match for their expectation. First, they have to go through sites which have reviews. It will also take time. And then he or she must find the reviews for their expectation. Sometimes even though they have gone through most of the comments they can't get the real idea whether that place is good or bad. If there is a system which gives summary about all these aspects and gives the recommendation, it will be easier for visitors. And, it will be more useful for service providers to find their weakness and improve quality of their services.

In most websites they provide rating system. But from that we can't get quantitative enough idea.

1.3 The exact computer science problem

In Past people took suggestion for their personal tourism from their friends or travel agencies. That traditional sources have serious limitations. When get a suggestion from a friend it is limit to the places that they have been visited. When get a suggestion from agencies they can be bias toward service they provide rather than suggest best for customer.

Because of prevalence of computer technology, now online sources are available for users to refer. It encourages users to use internet more and more. Because of that resource pool also growing so fast. Now there is a new problem. Referring and analyzing all these data manually is not easy. To get idea about something user have to refer most of available data. So, there should be an automated way to get summary picture of available data. There should be a way to suggest most relevant data user immediately without wasting user's time.

1.4 Objective

The main objective of this project is to extract useful specific features from review pool according to their demand and show them to user as summary. This summary will be created according to the aspect what travelers are expecting. Then travelers can decide their trip quickly and they don't need to spend much time on searching. This will also help travel service providers to improve quality of their service.

1.5 Scope

In here Author consider the review data related to hotel places in Ella, UVA province and store them in a database. Then these stored review data will be analyzed and categorized accordingly and rank them. Both positive and negative feedbacks are considered to get correct ranking. Author choose Ella town because there is considerable amount of review data related to hotel when comparing to other towns. Ella is one of most beautiful places in Sri Lanka and tourist tend to visit that place. Because of that reason there are considerable amount of review comments in TripAdvisor's site.

1.6 Summary

This chapter contains the problem that motivated author to find a solution. As explained in this chapter it will be very useful to have an automated system to read online hotel review comments. This will be useful to both service providers and the users. In this chapter, it also explained the computer science problem inside this problem, and it discussed the scope. User choose Ella district in UVA province as the scope. In next chapter user has explained the background study of related to this proposed approached.

Chapter 2

Literature Review and Background

2.1. Introduction

Recommendation systems are decision support tools when there is a large number of options. These systems are useful not only when users have a lot of options but also when they do not have domain-specific knowledge to make decisions. There are different kinds of recommendation systems and in this research, it is about a hotel recommendation system and it will be going to do by using aspect-based opinion mining.

When considering the extracting opinions from reviews there are several methods for doing this. Most approaches are based on natural language processing techniques and lexical resources and machine learning techniques.

Natural language processing and lexical resources: - Part of speech (PoS) identification and lexical databases are used in this approach. [2]

Machine learning: - Naïve Bayesian and Support Vector Machine (SVM) classification are used in this approach. [2]

- Naïve Bayesian method: - Using probability concepts and is based on Bayes theorem.
- Support Vector Machine: - Supervised learning method used for classification by recognizing patterns in data.

There are also opinion mining research methods that use multiple approaches combining supervised learning methods with lexical resources or ontologies, called hybrid approaches [2]

In next paragraph it will describe related works that was done by other researchers.

2.2. Related Works

Using Opinion Mining Techniques in Tourism

Bucur Cristian done research under title “Using Opinion Mining Techniques in Tourism”. In this research researcher proposes architecture content two module. [2]

- **Content acquisition module**

In this module it collects review data from website using web crawler and stored them in Review Deposit. Proposed solution uses MySQL database as storage

- **Analysis module**

In this module it will pre-process the data extract from content acquisition module and that will use for implements the opinion mining process. In opinion mining process it will process the text for each review and split it into sentences. Then review sentences are evaluate using POS tagging algorithm and the words polarity is evaluated using SentiWordNet.

This proposed platform was evaluated using a manually pre-classified dataset of review which was done by Enrique Vallés Balaguer and Paolo Ross researchers at the Natural Language Engineering (NLE) Lab, Universitat Politècnica de València. This corpus contains 3000 reviews and have been manually classified in positive and negative. [2]

Above mentioned research works are related this research. In this research also author have to collect data from websites and have to tokenize and get the polarity. As mention in this research pos tagging and SentiwordNet can be useful. These techniques and tool will be described in later chapters with details.

Mining Opinion Features in Customer Reviews

Mining Opinion Features in Customer Reviews is another research area done by university of Illinois. In here author targeted product sold online and going to analysis the review for these products using feature-based opinion mining. This approach has two steps. [3]

- **Feature extraction and opinion orientation identification**

Identify the features of the product that customers have expressed opinions on.

- Identify the positivity and negativity of each feature to decide review is positive or negative.

In this paper author only discussed about the first step. In here product name and entry page are used as inputs and summary of reviews is the output. When consider the review about the given product there are many ways to describe them. Some users use noun or noun phrase explicitly as product features. And some users use implicit features when they describe about the given product. In this research author only consider about the explicit feature. These explicit features will extract from review by using part-of-speech tagging. Then most frequent features will be extract using association mining. In next step system will prune the unnecessary features from generated frequent feature set because not all frequent features generated are useful or genuine features. Then system will extract the opinion features. After opinion features have been identified, Author determine the positivity or negativity of each opinion sentence using WordNet.

According to details mention in this research it is also a research which is extract opinion and get the positivity and negativity of each opinion using wordNet. In this research also author extract the opinion word and calculate the negativity and positivity.

Aspect based Opinion Mining from Restaurant Reviews

Aspect based Opinion Mining from Restaurant Reviews is another research project done by Indian government Engineering school. This paper focus on the aspect-based opinion mining. Because of growth of many social media applications, user now have many opportunities to express their review and idea through the internet. Individual and organizations use these published ideas for decision making. [10]

This paper aims to implement an aspect-based opinion miner for tourism domain, which automatically finds important features or aspects and its opinion. In here it will create a sentiment profile of each restaurant, which can be further used to compare and select restaurants at a particular location by a traveler. Author used 410 reviews for this research and average accuracy of this approach is 74%. Following are the steps they followed in their approach. [10]

- Collect data using web crawler and extract the review of restaurant.
- Preprocessing extracted data and remove unnecessary characters.
- Aspect extraction. In here they used Standard POS Tagger for tagging purpose
- Subject and object classification. In here subject sentence are going to identified and remove other sentences.
- Identifying aspect related opinion words.
- Scoring for pattern
- Finally aggregate the score of each aspect and produce an aspect-based summary

This research was done for restaurant review. According to author of that paper he also collected review from website and tokenized them, then extract the features and got the score for these features. He has done these for each and every restaurant separately and gave the aggregate summary for every restaurant. This study also little bit like the one that has gone to do in this research. In this research also it gets the review from website and tokenize them. But feature extraction is going to do for whole hotel set. Pos tagger is the tool that going to be used for tagged the word in review comments.

Aspect Based Sentiment Analysis to Extract Meticulous Opinion Value

There is another research under title “Aspect Based Sentiment Analysis to Extract Meticulous Opinion Value”. In this paper it mentioned about some theoretical details about the sentiment analysis. Sentiment analysis or opinion mining can be considered as a sub-problem under Natural Language Processing. In this paper it mentions two analysis process. [8]

- Document analysis- Analysis the whole document
- Sentence analysis- Analysis the sentence at a time

This paper introduces a concept of aspect value which tells how much clear or specific is the opinion that is being given using aspect tree. The objective of this research is to get an opinion value for a particular student. Various teachers give their remarks about a student and the algorithm analyses these remarks to return an opinion as well as aspect value using aspect tree. The aspect tree stores a defined set of aspects. The students are analyzed over these aspects. Such aspects can be student’s academics, sports, extra-curricular, co-curricular performance, personality traits etc.

Sentiment analysis of movie reviews

Sentiment analysis of movie reviews is another research area. This paper is targeting the movie reviews. Authors used SentiWordNet scheme to compute the document-level sentiment for each movie reviewed and compared the results with results obtained using Alchemy API.

In this paper authors have attempted to explore a new SentiWordNet based scheme for both document-level and aspect-level sentiment classification. This scheme locates the opinionated text around the desired aspect feature in a review and computes its sentiment orientation. For a movie, this is done for all the reviews. The sentiment scores on a particular aspect from all the reviews are then aggregated. Finally, a summarized sentiment profile of the movie on all aspects is presented in an easy to visualize and understandable pictorial form. [7]

As mention in the paper author has followed both document level and aspect level for review classification but he only mentioned aspect level method with more details.

According to the discussion in the related work, Opinion mining is a more advanced mechanism than other approaches since it is the technique that used by most authors. In opening mining aspect-based opinion mining is the more appropriate mechanism than document level and sentence level opinion mining. In aspect-based method, it digs in to words, gets its aspects and discusses about aspects. In the following section, the author explains the experiment setup.

2.3 Sentimental analysis (Opinion Mining)

Sentiment analysis, also referred to as opinion mining, is an approach to natural language processing (NLP) that identifies the emotional tone behind a body of text.

Sentiment analysis systems help organizations gather insights from unorganized and unstructured text that comes from online sources such as emails, blog posts, support tickets, web chats, social media channels, forums and comments

Sentimental analysis can be done at three levels. Followings are those three levels.

Document level: - Classify the review we got from a document and indicate that the whole document expressed a positive sentiment value or negative sentiment value. [1]

Sentence level: -Sentiment analysis is performed on the sentences rather than performed on the whole document. This level determines that each sentence shows a positive sentence level

sentiment, negative sentence level sentiment or neutral sentence level sentiment. In here two tasks are performed. [1]

- subjectivity classification.
- sentiment classification.

Feature /Aspect level: - feature level sentiment analysis is most appropriate level that suits for human opinion. It performs more accurate analysis. The aspect level classification provides summary of multiple reviews based on the feature-based opinion in aspect level sentiment analysis. [1]

2.4 Summary

In this chapter, it has described the related research works and techniques that researchers had followed. According to that works most of authors used Sentimental analysis as the technique to resolve their research problem. Sentimental analysis known as Opinion mining can be divided in to three parts. Document level, Sentimental Level and Aspect Level. In document level, it considers whole document and get the polarity for whole document. In sentence level it considers sentence by sentence and get the sentiment polarity, in aspect level it considers aspect wise and get the polarity for aspect. Most of studies has recommended Aspect level as the most suitable technique. In this research also Author going to resolve the problem using this technique.

In next chapter Author will describe the design and methodology for this research problem.

Chapter3

Design and Methodology

3.1. Introduction

This chapter contains the proposed solution and design for the research problem that addressed in this research. It has data collection, preprocessing, Aspect extraction, Opinion word extraction and Sentimental Score calculation. To collect data, author, use TripAdvisor web sites since it is the most popular among all over the world. These all steps will describe further, later on this chapter.

3.2. Research methodology

This research contains the process of data collection, Pre-processing, Aspect extraction, Opinion word extraction and Aspect based summary generation. Pre-processing improve the accuracy of process of opinion mining. Aspects extraction involves identifying the aspect which is most case noun or noun phrase. For this use NLP techniques. Next process is opinion word extraction. That involves Dependency parser. Not all word in sentence is an opinion word. So, must extract the correct opinion word based on aspect. Then Identifying the orientation of opinion word based on every aspect. In next step system will provide an aspect-based summary. Finally, system will provide the overall summary by aggregating all the aspect-based summary.

3.3. System Architecture

Figure 3.1 shows the overall architecture of the proposed solution. In here collected data is preprocessed and send it to Tokenize and Aspect extraction module and Dependency parser. Then extract the opinion word and send that to SentiWordNet to get sentiment score. Sentiment score is going to calculate aspect wise and hotel wise.

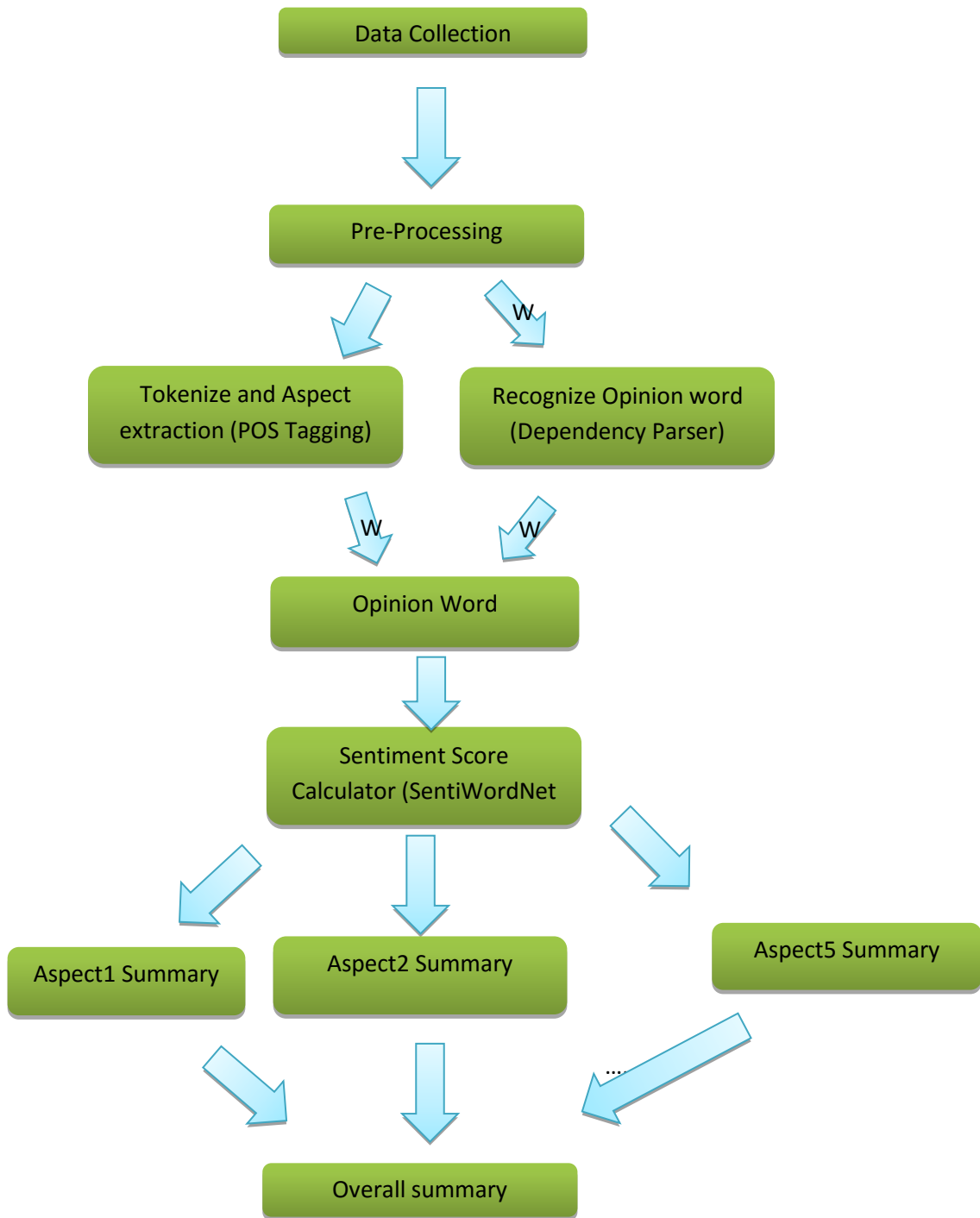


Figure 3.1 Proposed Design 0-1

Data will be collected from TripAdvisor website. Then the preprocessing will be done. In preprocess phase it will remove noise like “()”, Smileys and other unnecessary data. This preprocessed data will be sent through the Tokenize phase. In this phase we will use POS

tagging and we can recognize the Aspect here. Then opinion word will identify by using dependency parser. Both opinion words and aspect set will be sent to the SentiWordNet. Then sentences of the document would be evaluated. After that according to the opinion orientation, feature summary will be generated.

3.3.1. Data Collection

Data will be collected from TripAdvisor website. TripAdvisor [11] is the world largest travel site which contains approximately 661 million reviews and opinions covering the world's largest selection of travel listings worldwide. This site covering more than 200 hotel booking sites and covering approximately 7.7 million accommodations, airlines, experiences, and restaurants. In here author going to collect review comments related to hotels from TripAdvisor and save them into text database. Author only consider the latest five month's data since older data can be no longer valid.

3.3.2. Preprocessor

In sentiment analysis it will feed natural language. In this research it considers feedback which are published by hotel users as input. These feedbacks contain unnecessary characters and words, sometimes feedback contains malformed sentences, some use smileys instead of word to show their idea. Sometimes they use smiley to confirm their idea further with words. These kinds of unnecessary smileys, Characters like “!!!”, “()”, “#”, and unnecessary space can consider as noise in a sentence. If we fed sentiment analysis with these kinds of noise it will not work as expected way. It will not recognize or tokenize sentences correctly. This will affect the efficiency of the solution.

Before going to further steps, it is necessary to remove these kinds of noise from the feedback. In preprocessing phase, it will remove all unnecessary noise, and data will save to a text file database. Other than noise removing in this phase it will convert all sentences to simple case form. Because there is case sensitivity in SentiWordnet and it consider same word from capital and simple as two words. So, in preprocessing phase it will convert all database file to simple case. After preprocessed data, it will save to the text data files. These datafiles are going to be Inputs for Aspect extraction and Dependency parser.

3.3.3. Tokenize and Aspect extraction (POS Tagging)

When consider about review comments about any domain, they all are based on some aspects. Therefore, aspects are the most important fact regarding any domain's review comments. When consider hotel domain it plays a major role. Most of the time customer put comments against these aspects. In this research it consider aspect like rooms, staff, service ..etc. When people comment about hotels, they put them considering some aspect. As an example, if they want to tell something about foods in that place, they can put it like "food at that place is very good" or "Dinner is very good". So, in here considered aspect is food. People specifically can use word like breakfast, lunch or dinner. But they all refer food. In this research author consider as all word as food aspect.

First thing that going to be done in this research is tokenize review comment to find aspects. For that author use Pos tagger. In here preprocessed review fil are send to Pos tagger and it will tag all words in the review. Pos tagging happen as following.

POS (Part of Speech)

The part of speech [13] explain how word are place in a sentence. It explains which tag that a word can takes in a sentence. There are nine main tags that a word can get in a sentence.

1. Noun (N)- thing or person

Ex: - Daniel, London, table, dog, teacher, pen, city, happiness, hope

2. Verb (V)- action or state

Ex: - go, speak, run, eat, play, live, walk, have, like, are, is

3. Adjective(ADJ)- describes a noun

Ex: - big, happy, green, young, fun, crazy, three

4. Adverb(ADV)- describes a verb, adjective or adverb

Ex: - slowly, quietly, very, always, never, too, well, tomorrow

5. Preposition (P)- links a noun to another word

Ex: - at, on, in, from, with, near, between, about, under

6. Conjunction (CON)- joins clauses or sentences or words
Ex: - and, or, but, because, so, yet, unless, since, if
7. Pronoun (PRO)- replaces a noun
Ex: - I, you, we, they, he, she, it, me, us, them, him, her, this
8. Interjection (INT)- short exclamation, sometimes inserted into a sentence
Ex: - Ouch! Wow! Great! Help! Oh! Hey! Hi!
9. Determiner (DT)- limits or "determines" a noun
Ex: - a/an, the, 2, some, many

Beside these main tags there are several other tags. These tags can be found in the most popular tag set called Penn Treebank tag set. Most of the already trained taggers for English are trained on this tag set. Followings are few of them.

- NN Noun, singular or mass
- NNS Noun, plural
- NNP Proper noun, singular
- NNPS Proper noun, plural
- VB Verb, base form
- VBD Verb, past tense
- VBG Verb, gerund or present participle
- VBN Verb, past participle
- VBP Verb, non-3rd person singular present
- VBZ Verb, 3rd person singular present

In this section we only consider NN, NNS, NNP and NNPS.

A Part-Of-Speech Tagger (POS Tagger)

POS tagger is a process that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc., In here first thing happen is chunking the sentence. In chunking it is grouped every word in sentence to a group.

Following is the example which explain how pos tag happened.

Text need to be Pos tagged: - “food was brilliant, and the wine shipped in from nearby sister hotel appreciated”

Pos tagged output: -

food_NN was_VBD brilliant_JJ and_CC the_DT wine_NN shipped_VBN in_IN from_IN
nearby_JJ sister_NN hotel_NN appreciated_VBN

Aspect word get NN or NNP as tag. As mention in above theory they both refer noun or noun phrase. Most of the time aspect can be noun or noun phrase. So NN, NNS, NNP, NNPS are consider as aspect in this research. That means all the word which has tag staring with N is consider as aspects. To identify aspect, it considers all hotel file as one file and extract the aspect using whole file set. First noun and noun phrase are going to be identify and then going to identify the frequency of every aspect word. Then descending all of them into descending order and select top 5 words as aspect. Selecting to five is going to do manually. This to 5 are the aspect for this research.

3.3.4. Recognize Opinion word (Dependency Parser)

In here Author use Stanford dependency parser. It provides a simple description of the grammatical relationships in a sentence. This can be easily understood, and People can effectively use this without having linguistic expertise knowledge. In this parser it contains relations as type dependency relations. Dependencies are matched to directed graph representation.

Word in the sentence are nodes in the graph and grammatical relationship is shows as edge labels. These dependencies are triplets: Name of the relation, governor and dependent.

Here is an example sentence: “Bell, based in Los Angeles, makes and distributes electronic, computer and building products.” Figure 3.2 illustrates the graph for this example.

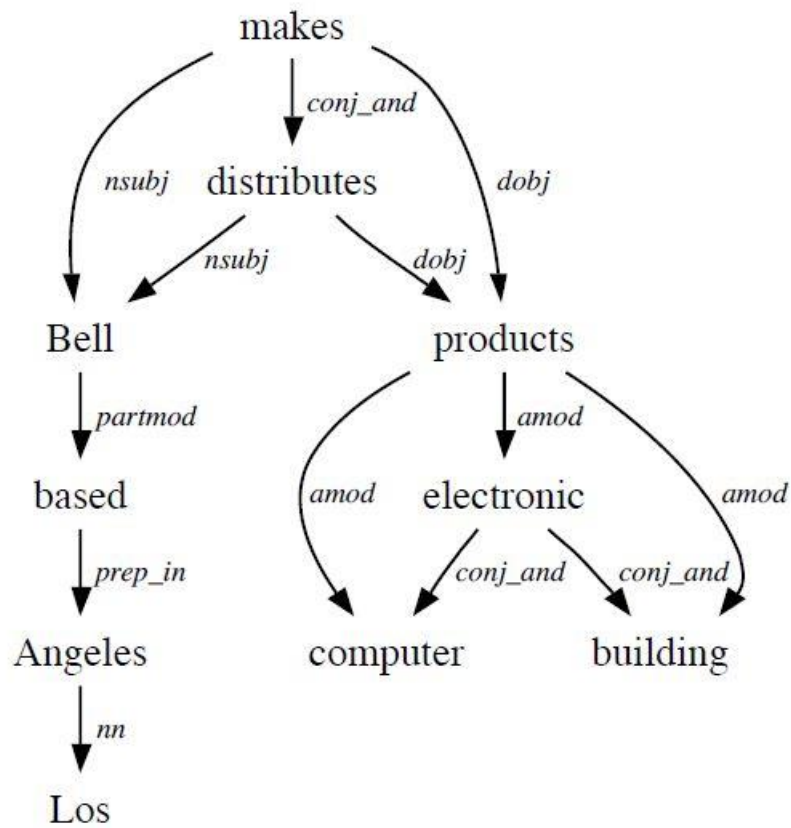


Figure 3.2 Dependency relationship 0-1

The subject of makes is Bell. It represents as following.

nsubj(makes-8, Bell-1)

The number in front of the word is the position of the word in the sentence and “nsubj” represents nominal subject.

In current Stanford representation contains approximately 50 grammatical relations. In this research author didn't concern about all 50 relations since most of them are not related to this research. This will discuss further, later in Implementation chapter.

3.3.5. Get Sentiment Score (SentiwordNet)

SentiWordNet [12] is a domain-independent publicly available lexical resource. There are around 117659 senti-synsets available in SentiWordNet. This number is very high compared to other lexicons.

The current "official" version of SentiWordNet is 3.0, which is based on WordNet 3.0. SentiWordNet is a text file and it contains 6 columns as POS, ID, PosScore, NegScore, SynsetTerms and Gloss. Table 3.1 contains the meaning of these columns. Figure 3.3 Is Screen shot of the SentiWordNet 3.0.

Column Name	Meaning
POS	part of speech which contains “adjective”, “verb”, “noun”, “adverb” (a,v,n,r)
ID	identification of the item
PosScore	Positive score; which has the weighted positive score value for a word
NegScore	Negative score; which has the weighted negative score value for a word.
SynsetTerms	reports the terms, with sense number, belonging to the synset.
Gloss	an explanation or definition of an obscure word in a text

Tabel 3.1 SentiwordNet description 0-1

```

# -----|
#
# POS    ID    PosScore    NegScore    SynsetTerms Gloss
a    00001740    0.125    0    able#1    (usually followed by `to') having the
a    00002098    0    0.75    unable#1    (usually followed by `to') not ha
a    00002312    0    0    dorsal#2 abaxial#1    facing away from the axis of
a    00002527    0    0    ventral#2 adaxial#1    nearest to or facing toward tl
a    00002730    0    0    acroscopic#1    facing or on the side toward the
a    00002843    0    0    basisopic#1    facing or on the side toward the l
a    00002956    0    0    abducting#1 abducent#1    especially of muscles; dra
a    00003131    0    0    adductive#1 adducting#1 adducent#1    especially of
n    15299225    0    0    study_hall#1    a period of time du
n    15299367    0    0    transfiguration_day#1    transfigurati
n    15299585    0    0    usance#1    the period of time perm
n    15299783    0    0    window#5    the time period that is
n    15300051    0    0    september_11#1 sept._11#1 sep_11#1
r    00001740    0    0    a_cappella#1    without musical acc
r    00001837    0    0    anno_domini#1 ad#1 a.d.#1    in the
r    00001981    0    0    common_era#1 ce#1 c.e.#1    of the
r    00002142    0    0    before_christ#1 bc#1 b.c.#1    before
r    00002296    0    0    bce#1 b.c.e.#1    of the period befor

```

Figure 3.3 SentiwordNet 0-1

In SentiWordNet it has several values for same opinion word. As an example, if we consider word “good” and if it is be an adjective it can be have several adjective values. Following Figure 3.4. shows that values. When we get a score for a word “good” SentiWordNet consider all those values and send weighted average as output.

```

a    0064787    0.625    0    good#5
a    00106020    0    0    good#2 full#6
a    00452883    0.5    0    near#5 good#10 dear#2
a    00523364    0.625    0    good#9
a    00775611    0.75    0    good#21

```

Figure 3.4 SentiWordNet value 0-1

In this research author going to feed output of dependency parser to SentiWordNet as key-value pairs. Before feeding this output to SentiWordNet, have to preprocess, Since we have to tell the SentiwordNet part of speech of the word which we are going to feed. This will further explain in implementation chapter. After we feed the opinion word with aspect SentiWordNet extract the opinion word, then it will get score value for opinion word. It will return the summarization of positive score and negative score. Summation value can be positive, negative or zero. If the value is positive that means opinion about aspect is good, if it is negative it means opinion about aspect is bad. If the summation is zero, opinion about aspect is neutral.

3.4. Problem Analysis

Problem that is going to be addressed through this research is getting hotel recommendations based on the reviews put by previous customers. Now a day's people use to check reviews before they are going to experience new things most of the time. This is quite common specially in hotel areas. These reviews are the most convenient way for tourists to get an idea about the place they are going to visit and stay. Not only tourists, local people and service providers also refer these reviews. But the problem is review count is too large to go through manually. So, people use to read top most reviews and get the idea. Most of the time that is not effective. Sometimes we get the idea that the place is good when we go through the first few reviews, but if we consider the overall comments it can be different. So, there should be an easy way to find this overall idea.

This research will try to provide a solution for the above mentioned problem. For that we must collect all the reviews regarding all the hotels in the defined scope. Most reviews have unnecessary data like smileys, brackets and unnecessary words. Before analyzing further, we should remove them first. Then we must identify the aspects mentioned in the reviews. Regarding hotel area rooms, food, service, location, staff are some aspects that can be considered by customers in the reviews. On some reviews all these aspects are not mentioned. Most reviews contain about rooms, food and service but not about the staff. Likewise, aspects are not distributed equally through reviews. This must also be addressed through this research.

Sometimes when all the review comments are considered overall idea can be positive, but reviews about staff can be negative. In this kind of scenario if there is one negative point it must be mentioned. Most of the time this kind of feedback can be useful to the service provider. With the reviews they can replace their staff, or they can train them to provide a better service. So, providing that as a message must be addressed as well. Although overall comments are positive, if there is any negative point that must be mentioned. Sometimes customer's main expectation can be the one which has a negative score. So besides providing recommendations if there is any negativity that must also be provided.

3.4 Summary

This chapter explained the design architecture of the proposed system. It described the all Steps that going to implement in next chapter. This design architecture contains Data collection, Preprocessing, Aspects extraction, Opinion word extraction and Sentiment Score calculation. For data collection it uses TripAdvisor websites since it is the most popular websites all around the world among travelers. In preprocessing phase, it removes unnecessary noise from the data. In aspect extraction phase it uses POS tagging technique and for Opinion word extraction it uses Dependency parser technique. For last phase sentiment score calculation, it uses Sent wordnet which is the extended version of Wordnet. In next chapter it will describe how author has implemented this proposed design using these techniques.

Chapter 4

Implementation

4.1. Introduction

In this chapter it contains detail description about the implementation of Aspect based sentimental analysis of Sri Lankan Hotels. It will explain tools and technologies that has been used in every process.

As mention in chapter 3 this research implementation also divided to 4 phases. First phase is preprocessing and in there it will remove all the noisy and unnecessary characters. Second phase is Tokenize and Aspect extraction. In here it will tokenize sentence using pos tagger and then extract the aspect related to hotel domain. Third is identifying opinion word. Finally, fourth one is getting Sentiment score. For that author used SentiWordNet package. In here it will find out the polarity of the aspect word.

4.2. Preprocessing

In this research it uses hotel reviews that was put by customers. Most of the time they are not well-formed sentences. Sometime in review comments it contains smileys and other unnecessary characters. Sometimes they use smiley to confirm their idea further with words. These kinds of unnecessary smileys, Characters like “!!!”, “()”, “#”, and unnecessary space can consider as noise in a sentence. If we fed sentiment analysis with these kinds of noise it will not work as expected way. It will not recognize or tokenize sentences correctly. This will affect the efficiency of the solution.

Before going to further steps, it is necessary to remove these kinds of noise from the feedback. In preprocessing phase, it will remove all unnecessary noise, and data will save to a text file database. Other than noise removing in this phase it will convert all sentences to simple case form. Because there is case sensitivity in SentiWordnet and it consider same word from capital and simple as two words. As example it will consider “GOOD” and “good” as two words. So, in preprocessing phase it will convert all database file to simple case. After preprocessed data, it will save to the text data files. These datafiles are going to be Inputs for Aspect extraction and Dependency parser.

Java is the technology that author use to do this task. No need to use special library here. Following is the pseudocode for this preprocessor.

Input the unprocessed text file

Read line by line

If noise found

Replace with ""

Write to a new text file

4.3. Tokenizing and Aspect extraction (POS tagging)

Preprocessed data base was sent to the tokenize and aspect extraction module. In here all files are consider as one file and feed that to this module. If we feed data to this module one by one it generates aspect regarding each hotel separately. But we want to get aspect for whole hotel industry, so had to consider whole hotel's comments as one file. Then it generated aspects for Hotel domain.

In here first thing has done was tokenize the whole file. For that Author used Part of Speech tag mechanism which we discuss in chapter 3. As mention in figure 4.1 it tokenized whole file set.

For this Author used "Stanford Log-linear Part-Of-Speech Tagger". This is a java tool that reads text in some language and assigns parts of speech to each word. Author used this library inside the program. This can be download from Stanford web site. The full tagger contains four trained taggers English, Arabic, Chinese and French. English tagger is the one that use in this research since most of comments are written in English. Following pre-requisites were needed to use this tagger. The system requires Java 1.8+ to be installed. Depending on whether you're running 32- or 64-bit Java and the complexity of the tagger model, you'll need somewhere between 60 and 200 MB of memory to run a trained tagger (i.e., you may need to give java an option like `java -mx200m`). In this research Author used tagger version 3.1.3. There are several new versions are available. But this is the one fully tested and stable. So, author used that in this research. Tried two latest versions, but it was

not successful. Some issue was occurred during tokenize process. Because of that use the 3.13 version.

Following figure 4.1 shows out put of Pos tag file.

```

Fantastic_JJ place_NN !. two_CD couples_NNS were_VBD here_RB for_IN 3_CD nights_NNS , and_CC all_DT of_IN us_PRP were_VBD very_RB satisfied_JJ with
t_VB to_TO leave_VB the_DT most_RBS perfect_JJ spot_NN in_IN the_DT area_NN ? . we_PRP d_NN arrived_VBD and_CC apart_RB from_IN a_DT bracing_VBG walk
obby_NN . . food_NN was_VBD brilliant_JJ and_CC the_DT wine_NN shipped_VBN in_IN from_IN nearby_JJ sister_NN hotel_NN appreciated_VBN . . we_PRP staye
_PRP$ team_NN running_VBG a_DT excellent_JJ kitchen_NN with_IN amazing_JJ food_NN , , if_IN you_PRP eat_VBP his_PRP$ food_NN once_RB will_MD defiantly
otherwise_RB it_PRP is_VBZ a_DT beautiful_JJ hotel_NN with_IN great_JJ food_NN . . just_RB the_DT most_RBS perfect_JJ hotel_NN we_PRP stayed_VBD in_IN
l_PDT the_DT amenities_NNS of_IN a_DT fully_RB equipped_VBN hotel_NN and_CC discreet_JJ service_NN staff_NN . . good_JJ wifi_NNS and_CC cable_NN tv_NN
. . nice_JJ infinity_NN pool_NN - : smallish_JJ but_CC sufficient_JJ and_CC we_PRP saw_VBD some_DT weasel_NN scurrying_VBG past_IN the_DT area_NN as_RE
ion_NN and_CC courtyard_NN could_MD do_VB with_IN a_DT bit_NN of_IN interior_JJ design_NN to_TO make_VB it_PRP more_RBR welcoming_VBG , but_CC we_PRP
out_RP to_TO be_VB one_CD of_IN the_DT best_JJS hotels_NNS -LRB- -LRB- out_IN of_IN 8_CD -RRB- -RRB- . . very_RB clean_JJ and_CC quiet_JJ . . we_PRP lc
h_WDT also_RB was_VBD not_RB a_DT problem_NN at_IN all_DT . . just_RB be_VB aware_JJ of_IN your_PRP$ shower_NN timings_NNS . . great_JJ location_NN ,
efinitely_RB stay_VB here_RB again_RB . . lovely_JJ pool_NN . . great_JJ view_NN . . great_JJ breakfast_NN . . good_JJ suppers_NNS . . ella_NN town_NN
uest_JJ friendly_JJ and_CC perfect_JJ for_IN a_DT weekend_NN getaway_NN . . i_LS would_MD suggest_VB it_PRP to_TO anyone_NN who_WP wants_VBZ to_TO have
ecial_JJ request_NN also_RB possible_JJ . . thank_VB you_PRP chef_NN prasad_NN and_CC the_DT kitchen_NN team_NN . . ex_FW butler_FW herath_NN , , chami
to_TO little_JJ adam_NN 's_POS peak_NN and_CC the_DT nine_CD arches_NNS bridge_NN -LRB- -LRB- which_WDT you_PRP can_MD also_RB see_VB from_IN the_DT gr
to_TO colombo_NN . . this_DT place_NN is_VBZ unlikely_JJ to_TO be_VB a_DT secret_NN for_IN very_RB long_JJ ! . we_PRP spent_VBD three_CD nights_NNS in_IN
onestly_RB . . staff_NN members_NNS were_VBD help_VB the_DT any_DT emergence_NN situation_NN in_IN our_PRP$ room_NN . . finally_RB we_PRP would_MD like
ool_NN . . only_RB stayed_VBD a_DT couple_NN of_IN nights_NNS , , but_CC could_MD easily_RB have_VB stayed_VBN longer_RBR . . a_DT 10_CD room_NN hotel
ts_NNS there_RB . . on_IN our_PRP$ last_JJ afternoon_NN , , we_PRP had_VBD a_DT light_JJ snack_NN on_IN the_DT lawn_NN watching_VBG the_DT sun_NN set_I
N lankan_NN breakfast_NN -LRB- -LRB- needs_NNS to_TO be_VB ordered_VBN the_DT day_NN before_IN -RRB- -RRB- which_WDT was_VBD well_RB prepared_VBN . .
to_TO relocate_VB across_IN . . unsure_JJ how_WRB painful_JJ or_CC inconvenient_JJ this_DT might_MD be_VB as_IN we_PRP did_VBD n't_RB use_VB this_DT se
P like_IN authentic_JJ places_NNS you_PRP can_MD find_VB very_RB comfortable_JJ and_CC quiet_JJ nice_JJ view_NN in_IN the_DT mountain_NN and_CC the_DT
_JJ after_IN a_DT busy_JJ day_NN although_IN the_DT water_NN very_RB chilly_JJ . . the_DT restaurant_NN was_VBD also_RB great_JJ , , we_PRP stayed_VBD
company_NN which_WDT is_VBZ definitely_RB worth_JJ a_DT visit_NN . .

```

Figure 4.1 PosTagger Output 0-1

Following is the example which explains how pos tag happened.

Text need to be Pos tagged: - “food was brilliant, and the wine shipped in from nearby sister hotel appreciated”

Pos tagged output: -

food_NN was_VBD brilliant_JJ and_CC the_DT wine_NN shipped_VBN in_IN from_IN
nearby_JJ sister_NN hotel_NN appreciated_VBN

In here pos tagger recognized each of word and it assigned part of speech to each word. According to this it recognize food, wine, sister, hotel are as noun (NN), Shipped and appreciated are as verb, past participle (VBN) , brilliant and near by as adjectives(JJ) , and “and” as coordinating conjunction (CC). Likewise, this pos tagger recognized the part of speech of every word.

Like above example every word in every hotel review file are tagged as shown in figure 4.1. Then author extract only the word which tagged as NN, NNS, NNP, NNPS since aspect can be take only one of these tags. Extracted word and its tags were written to another file like

```
place_NN
couples_NNS
nights_NNS
stay_NN
ella_NN
place_NN
traffic_NN
km_NN
center_NN
cabins_NNS
bedroom_NN
bathroom_NN
bathtub_NN
view_NN
balcony_NN
pool_NN
area_NN
infinity_NN
pool_NN
food_NN
service_NN
staff_NN
place_NN
food_NN
service_NN
heaven_NN
sentwe_NN
trips_NNS
ella_NN
earth_NN
spot_NN
area_NN
```

Figure 4.2 Extract Aspect Word List 0-1

shown in figure 4.2.

Then program calculate the number of occurrences of every word and it will also write to a file according to descending order. After that written into a file, top five aspect were extracted using human knowledge.

Followings are the outcome of the number of occurrences of every word. In here the author considers top 15 elements and selected 5 aspect with the help of human intelligence.

room_nn : 4611

ella_nn : 3929

hotel_nn : 3727

view_nn : 3697

place_nn : 3620

rooms_nns : 3189

food_nn : 3092

breakfast_nn : 2803

staff_nn : 2748

views_nns : 2603

sri_nn : 2251

tea_nn : 1918

stay_nn : 1876

restaurant_nn : 1863

night_nn : 1835

In here room and rooms are referred the same aspect, there for the full count for room aspect is 7800 (4611+3189). And considered “room” as the first aspect. Second top element is “ella” and we can neglect it using human knowledge. When consider hotel and place, author neglects that element since it is something which talks about the overall idea of the hotel. So, using human knowledge author removes that elements from the aspect list. Next one is “view” and here, view and views can be referred as the same thing and so that author considers “view” as an aspect. Next one is food and it is considered as an aspect. The other one is staff and it also can be considered as an important aspect with respect to the hotel domain. For the next author consider restaurant as aspect since other value can be neglected

using human knowledge. So, in this research room, food, staff, view and restaurant are the top most aspects that the author considers in further steps.

4.4. Dependency Parsing

Preprocessed database files were sent to dependency parser to get grammatical relationship between words in a sentence. After sent to dependency parser it tagged ever word in sentence according to there part of speech and then send the grammatical relationship as mention in the chapter3.

As an example, we sent the following sentence to dependency parser module and it gave output as following.

Sentence sent to dependency parsing module: -

“the food and the service from the staff were excellent! very recommended if you want a quiet and very nice place to relax, good food and service.”

Sentence were pos tagged as follows: -

the_DT food_NN and_CC the_DT service_NN from_IN the_DT staff_NN were_VBD
excellent_JJ ._.

very_JJ recommended_VBD if_IN you_PRP want_VBP a_DT quiet_JJ and_CC very_RB
nice_JJ place_NN to_TO relax_JJ ,_, good_JJ food_NN and_CC service_NN ._.

Universal dependencies for above sentences as following: -

det(food-2, the-1)

nsubj(excellent-10, food-2)

cc(food-2, and-3)

det(service-5, the-4)

conj(food-2, service-5)

case(staff-8, from-6)

det(staff-8, the-7)
nmod(service-5, staff-8)
cop(excellent-10, were-9)
root(ROOT-0, excellent-10)
nsubj(recommended-2, very-1)
root(ROOT-0, recommended-2)
mark(want-5, if-3)
nsubj(want-5, you-4)
advcl(recommended-2, want-5)
det(place-11, a-6)
amod(place-11, quiet-7)
cc(quiet-7, and-8)
advmod(nice-10, very-9)
conj(quiet-7, nice-10)
dobj(want-5, place-11)
case(food-16, to-12)
amod(food-16, relax-13)
amod(food-16, good-15)
nmod(place-11, food-16)
cc(food-16, and-17)
conj(food-16, service-18)

In above output if place is aspect then that dependency was saved to use them in future steps. So in here we saved amod(place-11, quiet-7). But we didn't save det (place-11, a-6), since det

mean “determiner” and It doesn’t explain about the aspect. We save every dependencies like amod(food-16, good-15) since amod means adjectival modifier. The adjectival modifiers(amod) are saved because it modifies the meaning of the noun phrase properly and some dependencies like nsubj(excellent-10, food-2) also saved for future steps. Final output of this module is opinion word and their belonging aspect. Then these aspect and opinion word are sent to the next module SentiWordNet to get sentiment score.

Java is the program language use here

4.5. Sentiment Score Calculator (SentiWordNet)

In this proposed method it sends output of dependency parser as key value pair to SentiwordNet. We should specifically mention that the opinion word is adjective, adverb, noun or verb before feeding it to the SentiWordNet. We can send it like the following “a_good”. That means opinion word “good” plays adjective role here and calculate the values considering it as adjective.

Hotel Name	Aspect	Opinion word	Sentiment score	Overall(positive/negative)	Overall Score
A	a	x	0.75	Positive	0.46
		y	-0.5		
	b	d	0.69	Positive	
		e	-0.48		
B	a	x	0.48	Negative	0.14
		z	-0.59		
	b	f	0.75	Positive	
		g	-0.5		
C	a	y	0.69	Positive	0.72
		q	-0.4		
	b	e	0.85	Positive	
		f	-0.42		

Table 4.1 Sentiment Score Calculation 0-1

In above Table 4.1, it contains data related to 3 hotels called A, B, C. considering all three hotels review comments it has identified 2 aspect as a and b. for every aspect there are opinion words. That opinion words are in column three. Respective sentiment scores are in column 4 and polarity is mention in the column5. In column 6 it contains overall score of every hotel. When consider the hotel A it is good for aspect a and b. It got positive score for both aspects. So, customer feedback regarding hotel A is positive. When consider the hotel B, it is not good for regarding aspect a. It gets negative value for aspect a. That means that hotel is not good regarding aspect a. if a is food, this means hotel B's foods are not good. When consider overall, hotel C is the best place to stay when comparing all three hotels. It got the highest positive score.

Considering overall score of the hotels, author ordered the hotels according to descending order of the positive value. Then user can get decision by looking at the hotel list since they know the upper most names are the best places to visit.

4.6 Summary

Implementation chapter explains how the proposed design in chapter 3 was implemented in the implementation stage. The proposed chapter has 4 phases such as “Preprocessing”, “Aspect Extraction”, “Opinion word Extraction” and “Sentimental Score Calculation”. In this chapter it has described how they were implemented. To implement this approach author has used Java, Stanford “POS tagger”, “Dependency parser” and Sentiwordnet. For SentiwordNet author used java libraries.

Accuracy of implemented proposed approach will be discuss in next chapter.

Chapter 5

Results and Evaluation

5.1. Introduction

This chapter is focusing the get accuracy of proposed approach. In this chapter it explains how evaluation process happened. This process contains two phases. One is manual and other one is related to proposed system. In here it will evaluate the whole system considering all the hotels and it will evaluate as aspect wise. Manual testing phase is divided to other two phases. In first phase it will evaluate every hotel. In second manual phase it will consider one hotel and it will evaluate aspect wise.

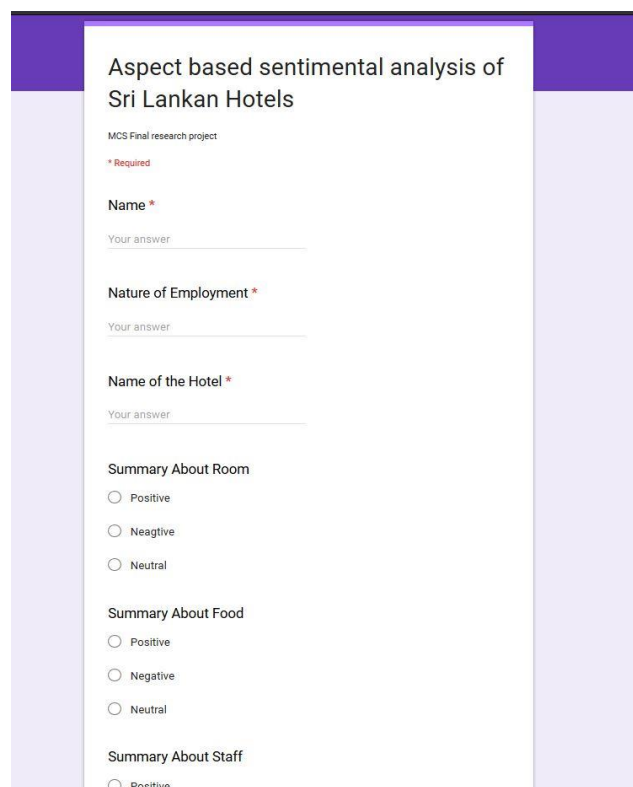
5.2. Manual Test phase

In this manual test phase author used two ways to do the testing and get the results.

Questioner is the technique used here and for that google form tool is used.

5.2.1 First manual phase

In here, author distributed data collected, regrading 20 hotels among 20 users. And provide the Questioner to them. They have gone through provided data set and they filled out the google form according to that.



The image shows a Google Form titled "Aspect based sentimental analysis of Sri Lankan Hotels". The form is for an "MCS Final research project" and includes a red asterisk indicating required fields. The form contains the following sections:

- Name ***: A text input field with the placeholder "Your answer".
- Nature of Employment ***: A text input field with the placeholder "Your answer".
- Name of the Hotel ***: A text input field with the placeholder "Your answer".
- Summary About Room**: Three radio button options: Positive, Neagtive, and Neutral.
- Summary About Food**: Three radio button options: Positive, Negative, and Neutral.
- Summary About Staff**: One radio button option: Positive.

Figure 5.1 Sample Google Form 1

Figure 5.1 is the appearance of the Google form that author provided to users. User provided their feedback aspect wise. Table 5.1 Illustrates the result extract from the first manual phase.

Name	Name of the Hotel	Summary About Room	Summary About Food	Summary About Staff	Summary About View	Summary About Restaurant
1	A	Positive	Positive	Positive	Positive	Neutral
2	B	Positive	Positive	Positive	Positive	Neutral
3	C	Positive	Positive	Positive	Positive	Positive
4	D	Neutral	Positive	Positive	Neutral	Positive
5	E	Positive	Positive	Positive	Positive	Positive
6	F	Negative	Negative	Positive	Positive	Positive
7	G	Neutral	Positive	Positive	Positive	Neutral
8	H	Neutral	Positive	Positive	Positive	Positive
9	I	Neutral	Positive	Neutral	Positive	Neutral
10	J	Neutral	Neutral	Positive	Positive	Neutral
11	K	Neutral	Positive	Positive	Positive	Neutral
12	L	Positive	Neutral	Neutral	Positive	Positive
13	M	Positive	Positive	Positive	Positive	Positive
14	N	Positive	Positive	Positive	Positive	Neutral
15	O	Positive	Positive	Positive	Positive	Neutral
16	P	Positive	Positive	Positive	Positive	Neutral
17	Q	Positive	Positive	Positive	Positive	Positive
18	R	Neutral	Positive	Positive	Positive	Neutral
19	S	Positive	Positive	Positive	Positive	Positive
20	T	Positive	Positive	Positive	Positive	Positive

Table 5.1 Result of manual Phase1 1

5.2.2 Second manual phase

In this phase, author distributed one dataset regarding to one hotel to 20 users and provided same questioners and collect results as the first manual phase. This phase results were used to calculate aspect level accuracy. Calculation can be done using following equation.

$$\text{Accuracy of aspect} = \left[\frac{(\text{number of matches with proposed system result})}{20} \right] * 100$$

Number of matches have selected considering the result which got from the proposed system. Got the count of records which exact match with the result got from the proposed system. Table 5.2 illustrate the result of hotel “T” which was produce by proposed system.

Name	Name of the Hotel	Summary About Room	Summary About Food	Summary About Staff	Summary About View	Summary About Restaurant
20	T	Positive	Positive	Positive	Positive	Positive

Table 5.2 Proposed System result 1

Table 5.3 illustrate the results set which was produce by second manual phase.

Name	Name of the Hotel	Summary About Room	Summary About Food	Summary About Staff	Summary About View	Summary About Restaurant
1	T	Positive	Positive	Positive	Positive	Neutral
2	T	Positive	Positive	Positive	Positive	Positive
3	T	Positive	Positive	Positive	Positive	Neutral
4	T	Positive	Positive	Positive	Positive	Neutral
5	T	Positive	Positive	Positive	Positive	Positive
6	T	Positive	Positive	Positive	Positive	Positive
7	T	Positive	Positive	Positive	Positive	Positive
8	T	Positive	Positive	Positive	Positive	Positive
9	T	Positive	Positive	Positive	Positive	Positive
10	T	Positive	Neutral	Positive	Positive	Positive
11	T	Positive	Positive	Positive	Positive	Positive
12	T	Positive	Positive	Positive	Positive	Positive
13	T	Positive	Positive	Positive	Positive	Positive
14	T	Positive	Positive	Positive	Positive	Positive
15	T	Positive	Positive	Positive	Positive	Positive
16	T	Positive	Positive	Positive	Positive	Positive
17	T	Positive	Neutral	Positive	Positive	Neutral
18	T	Positive	Positive	Positive	Positive	Positive
19	T	Positive	Positive	Positive	Positive	Positive
20	T	Positive	Positive	Positive	Positive	Positive

Table 5.3 Second Manual Phase Result 1

When consider the both Table 5.2 and Table 5.3, we can calculate aspect level accuracy as following example.

Aspect: Food

$$\text{Accuracy of aspect} = \left(\frac{18}{20}\right) * 100 = 90\%$$

5.3 Final Evaluation

Finally, it will calculate the final accuracy of the system by comparing both the result of Manual testing and the proposed system results. Author used following equation for that.

$$\text{Accuracy of Proposed System} = \left[\frac{(\text{Number of match in first manual})}{20} \right] * 100$$

	Name of the Hotel	Summary About Room	Summary About Food	Summary About Staff	Summary About View	Summary About Restaurant
1	A	Positive	Positive	Positive	Positive	Neutral
2	B	Positive	Positive	Positive	Positive	Neutral
3	C	Positive	Positive	Positive	Positive	Positive
4	D	Positive	Positive	Positive	Positive	Neutral
5	E	Positive	Positive	Positive	Positive	Positive
6	F	Negative	Negative	Positive	Positive	Positive
7	G	Positive	Positive	Positive	Positive	Neutral
8	H	Neutral	Positive	Positive	Positive	Positive
9	I	Positive	Positive	Positive	Positive	Positive
10	J	Positive	Positive	Positive	Positive	Neutral
11	K	Neutral	Positive	Positive	Positive	Neutral
12	L	Neutral	Positive	Positive	Positive	Positive
13	M	Positive	Positive	Positive	Positive	Neutral
14	N	Positive	Positive	Positive	Positive	Neutral
15	O	Positive	Positive	Positive	Positive	Neutral
16	P	Positive	Positive	Positive	Positive	Neutral
17	Q	Positive	Positive	Positive	Positive	Positive
18	R	Neutral	Positive	Positive	Positive	Neutral
19	S	Positive	Positive	Positive	Positive	Positive
20	T	Positive	Positive	Positive	Positive	Positive

Table 5.4 Proposed System Result 1

Table 5.4 illustrate the results of proposed system. When consider Table 5.1 and Table 5.4 ,14 records out of 20 are matched. Highlighted lines are the ones that are mismatch. According to that author calculate the accuracy of proposed system as the following.

$$Accuracy\ of\ Proposed\ System = \left[\left[\frac{(14)}{20} \right] * 100 \right] = 70\%$$

Figure 5.3 illustrate the summary of accuracy in a pie chart.

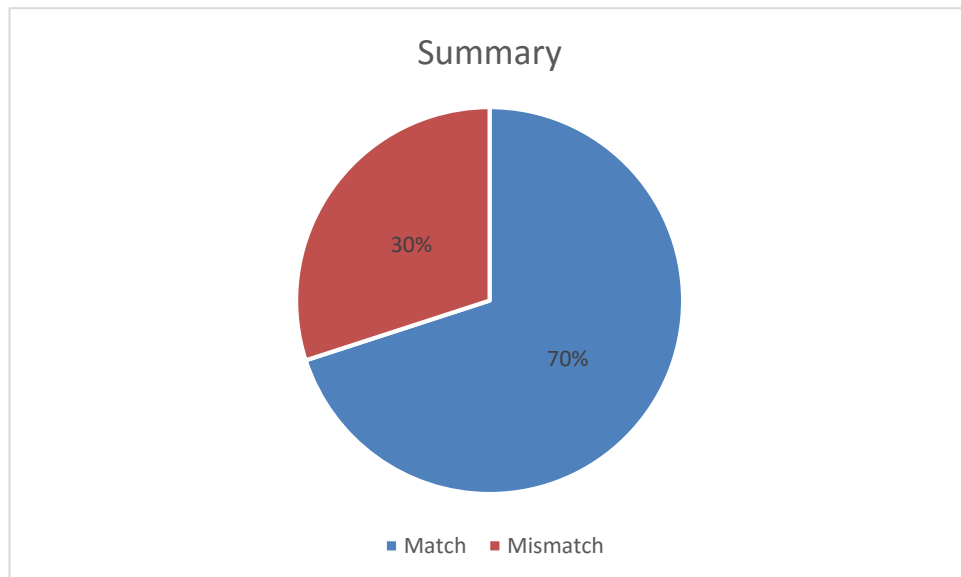


Figure 5.3 :Summary result of Accuracy 1

5.4. Summary

This chapter explained the results of proposed approach and how evaluation process happened. Author distributed review contains in TripAdvisor site to 20 members and got the result. 20 members got two different hotels. One hotel is same to everyone and other one is unique to everyone. Then author collect their opinion aspect wise. Finally, it compares with the result came out from the proposed approach.

In there, 30% mismatch and 70% match were found. So, Accuracy of this system is 70% and it is acceptable value.

Chapter 6

Conclusion and Future works

6.1. Conclusion

Most of the review comments provide by the customer are unstructured and not organized in a pre-defined manner. These texts are usually difficult, time-consuming and expensive to analyze, understand, and sort through. These kinds of complex analysis can be done by using Sentiment analysis (Opinion mining). It is also known as opinion mining. In sentiment analysis, unstructured information could be automatically transformed into structured data of public opinions. This is the most suitable one for mine the opinion of human. It is a type of text analysis that classifies the human generated text and makes decision by extracting and analyzing the text. It extracts the hidden knowledge from the unstructured texts exist in form of patterns and relationships. Opinion can be categorized as positive and negative, then measure the degree of positive or negative associated with the event.

There are three type of opinion mining techniques. Document level, Sentence level and Feature /Aspect level. From all these three, aspect level pinion mining is the best suitable method for this research.

Author collected data related to research from TripAdvisor website. Then preprocessed these data to remove noise from the data. After preprocess has done author sent these data to next module called aspect extraction module. Aspect extraction happen based on Pos tagging technique. Again, Preprocessed data sent to the dependency parser to get grammatical relationship of the word in sentence. Considering aspect extraction output and dependency parser output, system extracted the opinion words. After got the opinion word it sent to the sentiment score calculator as key-value pair. In here it used SentiWordNet to calculate the sentiment score.

When evaluating this implemented methodology, as per the chapter 5, “Testing - manually” and “Testing – Proposed system” methods are used. In there, it gets the accuracy considering whole hotel and considering aspect vise.

This research is very useful for both customers and service providers since it provides hotels feedback regarding each aspect. For service providers that they can get idea that they have to be improved and for user that they can make decision easily.

6.2 Future Work

The proposed model is based on aspect-oriented opinion mining and it can be further improved to give more accurate results. One of the main improvements would be to increase the data set considered. Currently the author has considered Ella area in UVA province only, but it is possible to consider all the districts data, for mining. This will directly have an impact on the output.

In the current model smileys are removed before processing due to complexity. But as smileys directly represent emotions, it is better to use these in predictions. Furthermore, current system considers only the comments in English language. There is a possibility to extract comments from other languages as well as an enhancement.

In the current proposed model 5 aspects are considered for predictions. Factors that are not currently referred like service, pool and other domain specific aspect can be taken to do predictions.

Current model uses data mining for classifications, but there is a possibility to use other mechanisms like neural networks, regression models or statistical models and can investigate the performance by comparing results.

References

- [1]. Maryam, B. (2017). Sentiment Analysis at Document Level. [online] Available at: https://www.researchgate.net/publication/320729882_Sentiment_Analysis_at_Document_Level [Accessed 29 Aug. 2018].
- [2]. Bucur, C. (2015). Using Opinion Mining Techniques in Tourism. *Procedia Economics and Finance*, 23, pp.1666-1673.
- [3]. Hu, M. and Liu, B. (2004). Mining and Summarizing Customer Reviews. [online] Available at: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.76.2378&rep=rep1&type=pdf> [Accessed 29 May 2019].
- [4]. Sharma, R., Nigam, S. and Jain, R. (2014). Mining of Product Reviews at Aspect Level. *International Journal in Foundations of Computer Science & Technology*, 4(3), pp.87-95.
- [5]. Hasan, S., Ukkusuri, S. and Zhan, X. (2016). Understanding Social Influence in Activity Location Choice and Lifestyle Patterns Using Geolocation Data from Social Media. *Frontiers in ICT*, 3.
- [6]. T C, C. and Joseph, S. (2015). A syntactic approach for aspect based opinion mining. *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*. [online] Available at: <https://ieeexplore.ieee.org/document/7050774> [Accessed 29 May 2019].
- [7]. Piryani, R., Gupta, V., Singh, V. and Ghose, U. (2014). A System for Aspect-level Sentiment Analysis of Movie Reviews. *International Journal of IT-based Social Welfare Promotion and Management*, 1(1), pp.1-8.
- [8]. Virmani, D., Malhotra, V. and Tyagi, R. (2014). *Aspect Based Sentiment Analysis to Extract Meticulous Opinion Value*. Bhagwan Parshuram Institute of Technology.
- [9]. Cao, L., Luo, J., Gallagher, A., Jin, X., Han, J. and S. Huang, T. (2010). A worldwide tourism recommendation system based on geotagged web photos. *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. [online] Available at: <https://ieeexplore.ieee.org/document/5495905> [Accessed 29 May 2019].
- [10]. T C, C. and Joseph, S. (2014). Aspect based Opinion Mining from Restaurant Reviews. *International Journal of Computer Applications*. [online] Available at: <https://pdfs.semanticscholar.org/07c0/9b3c412c4d9b887f29b57a2891d2414a120e.pdf> [Accessed 29 May 2019].
- [11]. MediaRoom. (2019). *Media Center*. [online] Available at: <https://tripadvisor.mediaroom.com/us-about-us> [Accessed 15 Nov. 2018].
- [12]. moalla, I. (n.d.). *CHAPTER_4_SENTIMENT_ANALYSIS*. [online] <https://www.academia.edu>. Available at: https://www.academia.edu/33127363/CHAPTER_4_SENTIMENT_ANALYSIS [Accessed 14 Nov. 2018].
- [13]. D'Souza, J. (2018). *Learning POS Tagging & Chunking in NLP*. [online] <https://medium.com/greyatom/>. Available at: <https://medium.com/greyatom/learning-pos-tagging-chunking-in-nlp-85f7f811a8cb> [Accessed 19 Oct. 2018].