“Bestaurentz” – Traveler Decision Support System Using Natural Language Processing

K.E.G.A.P Kadurugasyaya

2021

“Bestaurentz” – Traveler Decision Support System Using Natural Language Processing

A dissertation submitted for the Degree of Master of Computer Science

K.E.G.A.P Kadurugasyaya
University of Colombo School of Computing
2021

UCSC

DECLARATION

I hereby declare that the thesis is my original work, and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis. This thesis has also not been submitted for any degree in any university previously.

Student Name: K.E.G.A.P Kadurugasyaya

Registration Number: 2017/MCS/043

Index Number: 17440437

<u>30/11/2021</u>

Signature of the Student & Date

This is to certify that this thesis is based on the work of **<u>Mr. K.E.G.A.P Kadurugasyaya</u>** under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by,

Supervisor Name: Dr. D.A.S Athukorale

30/11/2021

Signature of the Supervisor & Date

I would like to dedicate this thesis to Dr. D.A.S. Athukorale

ACKNOWLEDGEMENTS

This project would no longer were possible without the support of any humans. I would like to thank to my supervisor, Dr. D.A.S. Athukorale, who study my numerous revisions, first rate feedback, excellent encouragement and to your endurance, guidance, and help. I have benefited significantly from your wealth of knowledge and meticulous enhancing. I am extraordinarily thankful that you took me on as a scholar and persisted to have confidence in me over time.

Many thanks to all the contributors, all the staff in University of Colombo, School of Computing, for the kind help throughout my MSc. Additionally, I amplify my way to all my colleagues at University of Colombo, School of Computing for the non-stop encouragement and support, time, advice and ethical assist.

Last, but not least, my warm and heartfelt thanks go to my family for the wonderful help and desire that they had given to me. Thank you all for the strength you gave me.

ABSTRACT

Picking a suitable hotel investigating the accessible information is one of the most complicated undertakings for travelers when planning their journeys. Around thousands of reviews with a range of information existing in social media web sites these days. With the growth of decision support system, explorers can assess the existing selections easily. Consequently, the travelers can make their pronouncements easily. Intention of this development is to develop traveler decision support system using online reviews. Deep learning techniques were used to develop this system and for Sentiment analysis of the hotel reviews. By this system, the emotions and the related information can be gathered from online reviews and capable of summarizing the final result. Nevertheless, this project was to label hotel reviews as positive, negative, or neutral and finally allow users to search them by keywords like food, cleanliness, etc.

This research was done using the hotel reviews about the Sri Lankan hotels. Reviews were collected from the booking.com website. This website is a very popular website among local and foreign travelers in recent years. Nearly 66000 reviews were used as the data set for this research. All these reviews were written in the English language.

In this project, a very basic machine learning algorithm which is Naive Bayes was used to build the classifier first. Due to the low accuracy of the classifier, wanted to use a new machine-learning algorithm to create the classifier. Considering the recent results of the NLP researches and the internal implementation, to create the sentiment analysis model, the convolutional neural network was used which is a class of artificial neural networks. This algorithm is most commonly used to analyze visual imageries but recently it has given promising results with text classification also.

In order to directly classify the hotel reviews as positive, neutral, and negative, an emotional feature extraction mechanism was used to create the CNN classifier. This feature contains the list of emotional categories that are widely used and accepted by the Word Emotion Association Lexicon. It extracts the frequencies of emotional categories from the given textual data. The significance of using this feature can be measured in terms of the concept that it is necessary to obtain emotions from the texts as they convey significant information to identify hatred speeches, overexcited texts, encouraging comments, or such strong emotions, that are useful to classify them on the emotional basis. Ultimately incorporating the emotion extraction feature with the classification model, accuracy could be improved as expectedly.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS iii

ABSTRACT iv

TABLE OF CONTENTS v

LIST OF FIGURES vii

LIST OF TABLES viii

LIST OF ABBREVIATIONS ix

INTRODUCTION 1

- 1.1. Motivation 2
- 1.2. Statement of the problem 2
- 1.3. Research Aims and Objectives 2
- 1.4. Scope 4
- 1.5. Research Contribution 4
- 1.6. Structure of the Thesis 5

LITERATURE REVIEW 7

- 2.1. Literature Review 7
- 2.2. Research Gap Identified 12

METHODOLOGY 13

- 3.1. Problem Analysis 13
- 3.2. Proposing Model/Design 14
- 3.3. Data Set 15
- 3.4. Data Pre-processing 16
  - 3.4.1. Tokenizing the Data 17
  - 3.4.2. Normalizing the data 18
  - 3.4.3. Noise Removing 18
- 3.5. Sentiment Analysis Model 19
- 3.6. Naïve Bayes Classifier 20
- 3.7. Neural Network Model 21
- 3.8. Neural Network Model With Proposed Features 25
  - 3.8.1. Emotional Categories 25

3.8.2. Sentiment Features ... 26
3.8.3. Visual Marker Features ... 27
3.8.4. Text Features ... 27
3.8.5. Model Training ... 28

EVALUATION AND RESULTS ... 29

4.1. Confusion Matrix ... 31
4.2. Accuracy ... 31
4.3. Precision ... 32
4.4. Recall/Sensitivity ... 32
4.5. F1 Score ... 32

REFERENCES ... 35

APPENDICES ... 38

LIST OF FIGURES

Figure 1: Structure of the thesis.. 6
Figure 2: High level diagram of proposed model.. 15
Figure 3: Sample review from Booking.com 16
Figure 4: Collected reviews.. 16
Figure 5: Data preprocessing.. 17
Figure 6: Noise removing.. 19
Figure 7: Noise removing from data... 19
Figure 8: Sentiment Analysis Model 20
Figure 9: Review Classification 21
Figure 10: Network Architecture.. 23
Figure 11: Model Training 28
Figure 12: Naive bayes results... 29
Figure 13: CNN results without additional features 29
Figure 14: CNN results with additional features 30
Figure 15: Accuracy of Naive Bayes model with two classes 30

LIST OF TABLES

Table 1: Confusion matrix without additional features 31
Table 2: Confusion matrix with additional features 31

LIST OF ABBREVIATIONS

NLP	Natural Language Processing
CNN	Convolutional Neural Network
KBR	Knowledge Base Repository
UDC	User Data Collector
CUI	Customized User Interface
IRE	Intelligent Recommendation Engine
RS	Recommender System
KNN	K-nearest Neighbors Algorithm
JSP	Java Server Page
CF	Collaborative Filtering
ANN	Artificial Neural Networks
GPU	Graphics Processing Unit
LSTM	Long Short-Term Memory
NLTK	Natural Language Toolkit
NRC	National Research Council
PV	Polarity Value
SS	Sentiment Score
SIM	Semantic Similarity
POS	Part of Speech

CHAPTER 1

INTRODUCTION

Most people in the world like to travel, as people crave new places, cities and countries. The purpose of visiting for most travelers has become more of a pleasure than a business venture.

Travel plays a huge role in today's style of life. People fond to travel not only in their own country, but also in other countries. The terminus can be natural such as the sea, waterways, lakes or highlands etc. Also, it can be a place that offers a variety of food, a variety of fashions, building styles, a different weather, lifestyle or a place with antique value. Or it can be a place where they can spend hours or days, relax or unwind, take a break from their tired life. Like camping or going out for a day to help reduce their stress in a busy life.

Most of the time people tend to choose a place to go when they look at a review, article, photos or place comment from another traveler that are mostly found on a social media site or google map. Currently there are many social networking sites such as Trip Advisor, Booking.com, a lonely planet etc. When people find an attractive place, they usually search it on Google and look at ratings and reviews of other travelers which is the best way to determine if it is the best place to visit before travelling.

A [1] study on "online travel reviews" conducted by the Intelligence Systems Laboratory, Texas A&M University, confirmed that 97.7% of people read online reviews of other travelers, by conducting a survey. They also found that people used online updates to search for accommodation, dining, going, time to go and what to do. So online reviews play a big role in getting going right now

To get a brief overview of a place or hotel or restaurant visited by people, people should spend a lot of time reading and analyzing reviews of many places. This process is time consuming, and this will reduce the desire to travel to the destination.

Sometimes when reviews are read, people will only read few random reviews from a huge collection of reviews and make decisions. The problem arises when all those few reviews are highly positive and rest of them are negative. In that case people will decide the travelling destination, a very bad place, thinking it is a good place to go depending on the selected reviews. For example, some hotels or restaurants may have a good profile where everyone's eyes will easily catch on and mean many benefits. The food or accommodation of those hotels or restaurants cannot be as good as they say. However, randomly selected reviews were positive, and most were negative. If so, once that hotel or restaurant has been selected, people may feel overwhelmed. Also, when we read the reviews, we can get a better idea than the explanation given alone.

As a solution to all of those difficulties and as a way to improve tourism the decision-making tool for travelers using online reviews is targeted at this project. Also, from this project, it will be strengthened to strengthen the division of reviews.

1.1. Motivation

As mentioned in the research problem section these days people find it very difficult to choose a suitable place to eat, visit and stay. In finding the right place, online reviews play a big role that contains people's ideas, hobbies and kindness. To get a general idea people should read all the reviews but not randomly as there are positive and negative reviews. But as a result of reading only a few comments many people get the incorrect idea because they may lose a lot of information and face a lot of difficulties. The motivation for this study is therefore to make the decisions of people and travellers better and making easier while defending their desire to travel by announcing a program that recites and investigate online reviews and ratings in a fast and efficient way while containing precise information.

1.2. Statement of the problem

This project is to develop a system that helps travelers to make decisions when choosing a hotel or restaurant. Often, people use online reviews and ratings to determine their locations. If so, people should read and check lot of number of reviews and ratings. This is a time-consuming and tedious task. Because of that, most travelers read a few reviews and decide where to go. This leads to major problems as it results going to a bad place, thinking it is good for a visit. In this case, the purpose of this project has become to develop a system that reads all user reviews and extract emotions separately from reviews and finally summarize the results into positive, neutral and negative categories. It is a traveler's solution to easily identify the right place.

1.3. Research Aims and Objectives

From this venture, a web application will be developed and designed to make it easier to decide on the best places to travel.

1.3.1 Aim

The main purpose of this project was to assist people in making decisions in choosing the best hotels for traveller using reviews and to create a comparison platform. From this, the time people spend analysis the reviews will be decreased, and it will make it easier to get a complete picture about the quality of the destination and people will be able to choose the best place (hotel / restaurant).

The detailed goals of this project are,

- Reduce time spent reading reviews
- Rate journey's end (hotel / restaurant) by reading all appraisals (not random reviews)
- Choosing the best place (hotel / restaurant) according to travellers (preferably with food, better accommodation)
- Promoting tourism

1.3.2 Objectives

To achieve these goals, some main objects have identified to be fulfilled by this project. From those the initial objectives of this project are:

- To make a method for categorizing the reviews according to the specification like, food and facilities like swimming pool, kids' activity area, service etc.
- to make a method to classify the categorized reviews into 3 forms: good, bad and neutral. So, for this to improve the accuracy of the classification has become an object of this project
- to make a method to summarize the classified review results, which has provided in the classification method, providing the easiest comparison.

In conclusion, with all these objectives, the main objective of this project has become to provide a system, which is a combination of all these methods and NLP technologies; that can improve the accuracy of classification and summarization.

So that, the objectives of this project have become, to make a method for categorizing the reviews according to the specification, to make a method to classify the categorized reviews and to make a method to summarize the classified review results. And the specific objects of this project are to increase the accuracy of the classification and to increase the accuracy of the summarization.

1.4. Scope

The project includes about implementing a decision-making System; providing the most relevant travel destinations based on tourist reviews on nominated web sites. This program will recap reviews in areas under a few categories; It will be very helpful for novel travelers to get a good idea of the places Perrier for leaving, deprived of spending more time reading all the user reviews about that destination.

For this system, travel destinations will be selected only from Sri Lanka and the training dataset will be also created using the reviews which are belongs to Sri Lankan hotels. Moreover, for now, only the reviews written from English language will be selected. In addition to that the review classification is done into three categories; positive, negative and neutral, as it is more effective, efficient and easy to read and understand. For this project, natural language processing technologies are using.

1.5. Research Contribution

By this project, a web system will be developed with the purpose of making it easier to decide on the best hotels or restaurants to go and helping people to make decisions in choosing the best places for traveling. For this traveller reviews and emotions in reviews will be extracted and analysed. So, from this project, the time people spend reading traveller reviews from various web sites or social medias will be decreased and make it easier to get a brief idea about the quality of the destination and people will be able to choose the best place to travel. And this study will help to decrease the time spent in studying reviews, help to rate the place by reading reviews completely (not arbitrary reviews), help choosing the best destination conferring to travellers wish (finest food, finest accommodation) and will help boost tourism.

For that from this project, methods will be created for categorizing the reviews according to the specification like, food and facilities like swimming pool, kids' activity area, service etc., to classify the categorized reviews into 3 forms; good, bad and neutral and to summarize the classified review results, which has provided in the classification method, providing the easiest comparison. And for improving the accuracy emotion feature extractions will be used. In conclusion, from this project, a system will be developed, which is a combination of all these methods and NLP technologies; that can improve the accuracy of classification and summarization. Moreover, from this project, a data set will be created indexing the reviews which are relevant to the Sri Lankan hotels.

1.6. Structure of the Thesis

This research describes a research conducted to build up a system which can be categorized and summarise reviews in to three categories to help people in decision making. This paper includes five main sectors, introduction, literature review, methodology, evaluation and results and finally references. In the introduction chapter it describes the importance of this research; how it is important to society and the research contribution to the computer society. Also, from the introduction chapter aims and goals of this research and the scope has provided.

In the Literature review section, research and projects conducted in the same research area has provided. In this background study, the technologies and the accuracy of research findings are stated. Research gap identified from this project will also include in this chapter.

The problem analysis and how the clarification and summarization has handled in the project is discussed in the methodology section. It indicates the steps, technologies, developing process and neural network model with proposed features. The research findings and the overall results with future works are described in the evaluation and results. There, the calculated accuracy of the presented system and future works to improve the current analysis and summarization in order to improve the accuracy has discussed. In the references, the original resources that used to borrow ideas has stated.

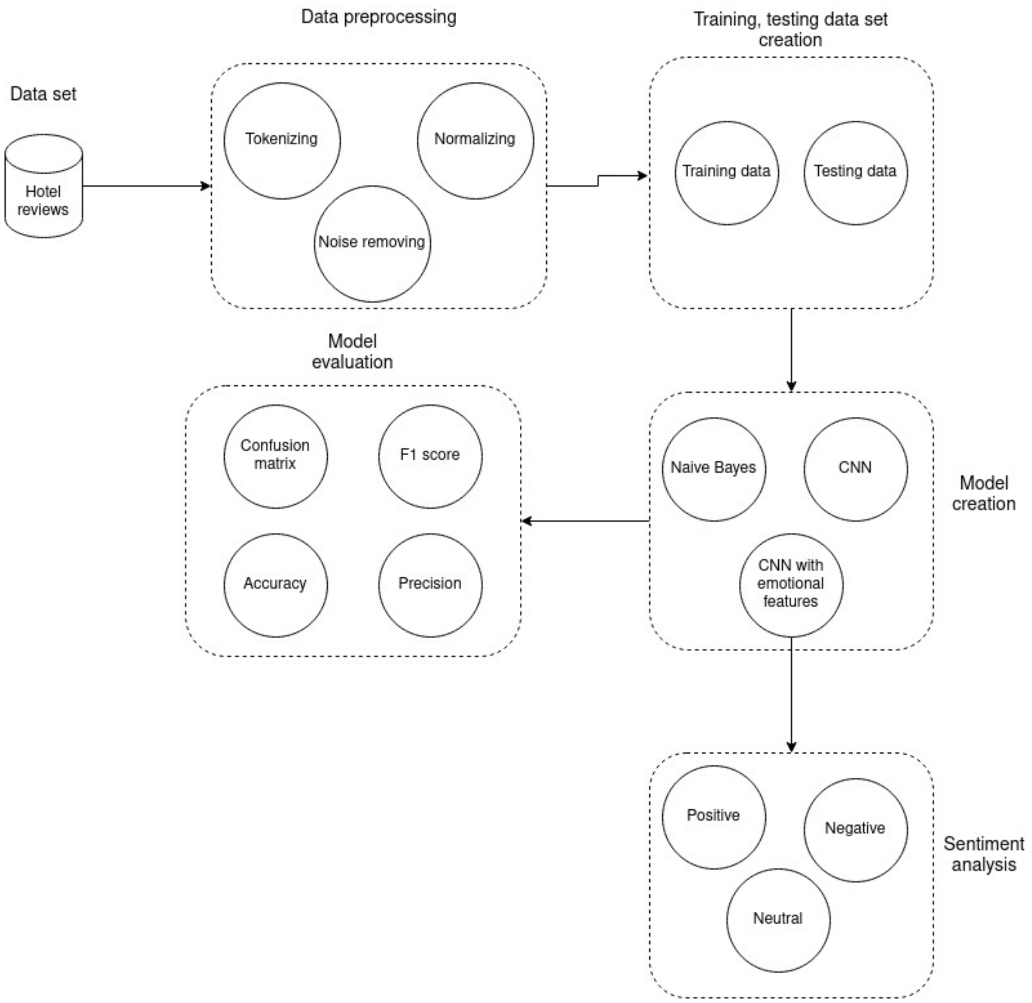

Figure 1: Structure of the thesis

Project data was collected from the booking.com web site. All of them were reviews about Sri Lankan hotels written in English. In the next step, data was preprocessed using techniques like normalization and tokenizing. Noise removing also had been done in this step, because data contained many numbers of unwanted values. In the next step, data set was divided as training and testing data sets with the ratio of 7:3. Next sentiment analysis models were created using Naïve Bayes and Convolutional Neural Networks (CNN) algorithms. One another model was created using CNN algorithm with embedding the text emotional feature extraction mechanism. Finally, these models were evaluated using the parameters like confusion matrix, F1 score, accuracy, and precision. Ultimately identified the best sentiment analysis model for Sri Lankan hotel reviews.

CHAPTER 2

LITERATURE REVIEW

There are a number of studies have been achieved for the accomplishing of traveler decision support systems in travelling circumstances. As quoted in the previous segment and bestowing to the study (Gretzel, Yoo & Purifoy 2007) regarding “online travel review” accomplished by Laboratory for intelligent systems in tourism, Texas A&M University, conducted a survey and has evidenced that 97.7 percent of people review other travelers’ online reviews. They discovered that people use online reviews to search a place as accommodation, for dine-in, to visit, when to plan and what to accomplish. Online reviews play a major character in exploration in existent. When examining for answers to this glitch, past to present, many investigators and inventers tried to come up with an elucidation for this problem.

2.1. Literature Review

A tourism recommender system (Loh, Lorenzi, Saldana & Licthnow 2003) proposed for travel mediators using association and text analysis. A study (Bigdeli & Bahmani 2008) done in September 2008, suggested their system to Equivalence accuracy of cosine-based similarity and correlation-based similarity algorithms in tourism recommender systems. They appraised the procedures based on the data composed from a big travel organization in Tehran.
In 2010, a research group (Kenteris, Gavalas & Mpitziopoulos 2010) established a mobile tourism recommender system manipulating Collaborative filtering techniques. The system was skilled of catching context-aware user appraisals and rating for recommendations job.
A tourism recommender system framework called “Personalized e-government services” (Al-hassan, H. Lu & J. Lu 2010) which is an ontology-based recommender system was developed established on Knowledge Base Repository (KBR), User Data Collector (UDC), Customized User Interface (CUI) and Intelligent Recommendation Engine (IRE). And then in November of 2015, an ontology-based tourism recommender system grounded on spreading activation model (Bahramiana & Abbaspoura 2015) has developed by Z. Bahramiana and R. Ali Abbaspoura. In this a recommender system (RS) evaluates the prodigious number of POIs and delivers individualised recommendations to operators based on their favourites. And a content-based recommendation system was suggested, which uses the data about the user’s partialities and POIs and computes a degree of likeness between them. It chooses POIs, which have uppermost similarity with the user’s preferences. The suggested content-based recommender system was

heightened using the ontological information about tourism domain to characterise both the user profile and the recommendable POIs. The planned ontology-based recommendation procedure is accomplished in three steps together with: ontology-based content analyser, ontology-based profile learner and ontology-based filtering component. Operator's feedback bends the user's preferences using Spreading Activation (SA) strategy. It shows the planned recommender system is efficient and progresses the overall performance of the long-established content-based recommender systems. Afterward in 2016 another research crew developed a tourism recommender system (Chu, H. Wang, Zheng & Z. Wang & Tan 2016) and it was developed built on a hybrid approach using association rules and ontology.

In the similar year a set of researchers from Tokyo has developed a tourism recommendation system (Namahoot, Panawong & Brückner 2016). The recommended system was developed using KNN algorithms and semantic web rule language for Thailand. They used Java Server Page (JSP) for application of the system. In that year, a tourism terminus recommender system (Zheng, Luo, Xu, Yu & Lu 2016) has developed to answer cold-start problem of operator and item-based. In this system, opinion-mining technology is used to recognize the user predilections and applicable items/products. These essentials are attached into a hybrid collaborative filtering process by joining user and item based collaborative filtering approaches. For this artificial interactive module is used to dodge the cold start dilemma. in 2017 a system (Kashevnik, Ponomarev & Smirnov 2017) was developed with a multi-model approach and architecture, called "multi model context-aware tourism recommendation service".

Some studies have also been accomplished in case of multi-criteria recommender systems in tourism domain. A service (Li, Wang & Geng 2008) was established Improving customized services in mobile trade by a modern multi criteria rating tactic, which is tensor decomposition technique in developing a recommendation system built on multi-criteria CF. in 2011, three multi-criteria recommendation methods were developed by a group (Liu, Mehandjiev & Xu 2011) using clustering techniques to gather the data earlier to prediction and recommendation tasks. Correspondingly in 2018, research group (Nilashia, Ibrahima, Yadegaridehkordib, Samadc, Akbari & Alizadehfa 2018) attempted to come up with an explanation which could find best destination in decision making, titled as "Travelers decision making using online review in social network sites (A case on Trip Advisor)". But it was only for hotels and restaurants. They have had accomplished a research on developing a novel recommendation technique for hotel recommendations in e-tourism platforms using the multi-criteria ratings. They have used supervised machine learning techniques and unsupervised machine learning techniques to study customers' online reviews.

A research team (Anania & Paolo 2016) has projected an ontology-based custom-made recommendation for tourism and spare time activities. One more research team (Yordanova & Kabakchieva 2017) has planned a way to Sentiment Classification of Hotel Reviews in Social Media with Decision Tree Learning.

Opinion mining is a topical area of interest for Natural Language investigators. Peoples are approaching to develop systems that can be identified and organize opinion characterized in each text. A different research team (Dalal, Mukesh & Zaveri) has proposed a system for opinion summarization and classification of online product reviews applying semi supervised learning.

A research team (Tan, Wang & Xu) has done a sentiment analysis for product reviews in amazon. they have used traditional machine learning algorithms including Naïve Bayes, Support Vector Machines, K-nearest neighbor, Recurrent Neural Network.

A tourism recommender system using collaboration and text analysis (Loh, Lorenzi, Saldana & Licthnow) has developed by a group of researchers including Stanley Loh, Fabiana Lorenzi, Ramiro Saldaña and Daniel Licthnow to support travel agents in discovering tourist choices for clients. The goal of this system has reached through recommendations of towns and their tourist attractions. To do that, the system examined textual messages sent by operators (a customer and a travel agent) when cooperating in a remote Web chat, and it identifies consumer's interest areas according to a tourism ontology predefined in the system. The chat is specially created for this system and it was not open to nonregistered operators (only two people interacting at a time). This system was content based for the reason that the preferences of the customer were matched alongside the content of items in the database. Preferences are identified in the sent messages and were defined as interesting areas. The set of possible areas was predefined in a tourism ontology (created previously). But the system was a support system because it does not make recommendations straight to the customer. But it helped the travel agent, suggesting options. This system was focused only on the problem of finding cities and attractions for the customer and it was difficult for recommendation if a customer come up with no plan.

A research conducted by two researchers from Department of Computer Science KAIST Daejeon, Korea (Jo & Oh) proposed dual models: Sentence-LDA (SLDA; a probabilistic generative model, which assumes all words in a single sentence are generated from one aspect) and Aspect and Sentiment Unification Model (ASUM). SLDA and ASUM model the generative process of reviews. Based on the surveillance above, SLDA and ASUM constrain that all words in a single sentence be generated from one subject. ASUM is an extension of SLDA into which sentiment is incorporated. In ASUM, the words in a sentence are produced from the same pair of aspect and sentiment, which is call as "senti-aspect".

When studying on NLP in opinion mining, a group of scientists from Korea and India has develop a system (Bhattacharyya, Biswas & Kim) that can identify and classify opinion or sentiment as characterized in an electronic text. The extent of the research was to develop a system which can recognize particular sentence from a document and successively recognize the polar phrases between the sentences either positive or negative.

As people states their opinions in the web about products and services which they have used, it has become important to develop approaches of (semi-)automatically classifying and evaluating them. Dongjoo Lee[1], Ok-Ran Jeong[2] and Sang-goo Lee[1] from [1] School of Computer Science and Engineering, Seoul National University Seoul 151-742, Republic of Korea and [2]Department of Computer Science, University of Illinois at Urbana-Champaign Urbana, IL, 61801, USA; survey and analyze various techniques which have been developed for the main tasks of opinion mining. Constructed on the survey and analysis of the techniques, the research group provided an overall inclusive visualize of what should involve in developing a software system for opinion mining and had examined the dual tasks that are detailed to opinion mining: development of verbal resources and sentiment classification. In addition to that they handed over an opinion summarization by looking into the existing opinion mining systems, which extract opinion expression from bulky reviews and show in what manner individually system applies the methods in order to meritoriously summarize and stage the opinions.

To issues such as quickly produce great superiority recommendations even for gigantic data sets, a group of researchers (Almazro, Shahatah, Albdulkarim, Kherees, Martinez & Nzoukou 2010) has explored quite a few collaborative filtering techniques for instance, the item-based approach, which detect relationship between items and circuitously calculate recommendations for operators based on the relationships and another to that, they analyzed different algorithms of operator based and item-based techniques for recommendation generation. Their research on recommender system was mainly absorbed on finding ways to increase the performance, scalability or accuracy of the algorithms. Thus, Hybrid algorithms that combine countenances of user-based and item-based algorithms have been set up and other approaches using Rough Set Prediction, Slope One Scheme Smoothing and also another approach to build item-based and user-based algorithms.

A mobile tourism recommender system (Kenteris, Gavalas & Mpitziopoulos) by a group of Department of Cultural Technology and Communication University of the Aegean Mytilene, Greece, was developed to overcome prevailing systems collapse to utilize information, behaviors, ideas, evaluations, assessments, ratings by other tourists with comparable interests, which provide pitch for the cooperative production of tourist contented and travel recommendations. So, they extend this perception of travel recommender systems utilizing collaborative filtering techniques for descending improved recommendations and proposed the use of Wireless Sensor Network (WSN) inductions around tourist sites for providing mobile operators an appropriate and low-cost means for uploading tourist information and ratings about Points of Interest (POI) through their mobile devices. User ratings uploaded via WSN infrastructures are weighted higher to distinguish between operators that rate POIs using the mobile tourist guide application in straight proximity of the POI and others using the web away from the POI.

A German research group (Jannach, Karakaya & Gedikli 2012) has conducted a research on "Accuracy Improvements for Multi-criteria Recommender Systems" and staged some methods to leverage information derived from multi-dimensional ratings to improve the extrapolative accuracy of such multi-criteria recommender systems. In addition to that, they anticipated to use Support Vector regression to determine the relative importance of the individual criteria ratings and suggested combining user and item-based reversion models in a subjective approach. Beside the automatic adjustment and optimization of the grouping weights, different feature selection approaches have used to further improvements of the superiority of the recommendations.

A group of researchers (Nilashi, Jannach, Ibrahim & Ithnin 2014) who has studied on "Clustering- and regression-based multi-criteria collaborative filtering with incremental updates", implemented a new way and different algorithmic approaches to transact with situations that are actually quite mutual in faithful settings. This group have talked the data sparseness problem by clustering the data. Specifically, proposed to use an Ant System-based Clustering Algorithm (ASCA) and an Ant K-means algorithm (AK) to detect operator segments with parallel tastes. These clusters used for more dependable regression functions. To transact with the noise in the data predicament, Principal Component Analysis (PCA) applied and in so doing acknowledged the most imperative superiority magnitudes for the different operator segments. And to be able to incorporate the continuous stream of novel rating data which can arrive in online platforms, data processing chain boosts incremental updates using incremental SVR and PCA techniques. These approaches were used to immediately utilize newly obtainable information and constantly apprise the recommendation models. And their test outcomes showed that the clustering and noise subtraction techniques helped to improve the extrapolation accuracy by more than ten percent in all tested scenarios in terms of the RMSE and the MAE.

In most of the studies discussed in this literature review section, the accuracy is low. Another major area that missed in most of stated researches were Emotional categories. When extracting the details of people's comments is the emotions included in the comment. Emotions has lot more details in conclusion which directly express what that comment include. From this research include both text extraction and emotion extraction with improved accuracy of the system.

2.2. Research Gap Identified

A system will be implemented from this project, which is able to classify particular hotel reviews as positive, negative, neutral and finally make summarization on each classified reviews. For this, natural language processing and deep learning techniques are going to be used to text classifications and sentiment analysis. The data sets will be created using the reviews that are belong to Sri Lankan hotels will be collected from several social media sites, and only the reviews written from English language. Then, system divides the reviews into subcategories (E.g., food, price, accommodations etc.) and the reviews will be classified using the natural language processing, neural networks and sentiment analyzing techniques, in to three forms: positive, negative and neutral. For improving the accuracy of the classification emotion feature extractions are going to be used.

Moreover, the system will allow the user to search hotels and will give a big picture of summarization under a table, which contains the subcategories and the summarization results making it easier to choose the best hotel according to the traveler's desire.

In the above-related research works, there are many types of traveler decision support systems had been implemented. Most of those systems classify user reviews as positive and negative only and give some decisions depending on that classification. This research mainly focused on sentiment analysis on Sri Lankan hotel reviews. All the training and testing data sets were collected from the booking.com website and all these reviews were about the Sri Lankan hotels which were written in the English language. In this research, sentiment analysis was performed with three classification classes. They are positive, negative, and neutral as it is more effective, efficient and easy to read and understand.

To build the sentiment analysis model, deep learning techniques and natural language processing techniques were used. a neural network algorithm was used. It is convolutional neural network (CNN) algorithm. This algorithm is most used to analyze visual imageries but recently it has given promising results with text classification also.

Instead of directly applying the CNN for sentiment analysis model creation, it was incorporated with the text emotion extraction feature. Ultimately improved the model accuracy using this feature.

CHAPTER 3

METHODOLOGY

3.1. Problem Analysis

Considering the researches that have done on this topic in the literature review section, most of them are decision-making programs. Those systems were able to read the reviews and make some decisions. The problem is that if those decisions are not in line with the desire of the travellers', in that case those decisions are not worth it. The system which has developed from this project ultimately produces a summary report using user updates and it is capable of capturing the emotions which has included in user reviews. From this system, the travellers can match his needs and decide the most suitable and preferred destination.

These days 90 percent of travellers' reserve hotels using popular online booking systems such as Booking.com or TripAdvisor or Agoda. When a traveller search through a site, for a place or a searching area, as the results it will display many hotels and restaurants featured in several user reviews. The user must go through all the reviews and need to find the best matching place. The proposed system in this report has provided the best solutions to all these issues.

For that, the system implemented in this project will be able to classify hotel reviews as positive, negative, neutral and finally make summarization on each classified reviews.

The initial step, natural language processing and deep learning techniques are going to be used to text classifications and sentiment analysis. Data sets that need to train the models are going to be created using the hotel reviews of Sri Lanka by this project itself. These reviews are supposed to be collected from the social media sites.

Then, system divides the reviews into subcategories (E.g., food, price, accommodations etc.) and the results of three summarizations (positive, neutral and negative) will be shown under each subcategory.

Moreover, the system will allow the user to search and compare several hotels at the same time and will give a big picture of summarization under a table, which contains the subcategories and the summarization results.

In deep, the reviews that are belong to Sri Lankan hotels will be collected from several social media sites, and only the reviews written from English language will be collected. The training data set will be created using the collected reviews. Then the reviews will be categorized according to the specification like, food and facilities like swimming pool, kids' activity area, service etc. As the second step, classification will be done using the natural language processing and sentiment analysing techniques. From this step, the reviews will be classified in to 3 forms: positive, negative and neutral. This step has few sub steps like gather data, from internal data like surveys, customer service and customer relationship management (CRM) software etc. and external data like social media, which will be used in this project. This external data from social media websites will be collected using technologies like web scraping tool/framework or using APIs. Then Aspect-Based sentiment analysis will be performed. There data with opinion units will be pre-processed and then sentiment analysis model will be created, trained and tested. After that, an aspect or topic classifier will be created, trained and tested again. For the final step, aspect-based sentiment analysis results will be visualized. At last, the classified review results will be summarized providing the easiest comparison. Finally, a system, that can improve the accuracy of classification and summarization will be created.

3.2. Proposing Model/Design

Hotel reviews were scrapped from the booking.com website. Scrapped hotel reviews contained review text and review score (rating). In the next step, cleaned the collected data set and prepared it to train the model. To train the model, few machine learning algorithms and one neural network algorithm were used initially. Comparing the accuracy of each model, model with the highest accuracy was used to make the predictions. Predicted reviews from the selected model were stored in the solr search engine. Solr engine is an opensource search platform written in Java. Full text search, text highlighting and the real time indexing are the main features of it. Finally classified hotel reviews are going to be displayed using a Java EE web application. This web application allows us to filter reviews by hotels and some keywords.

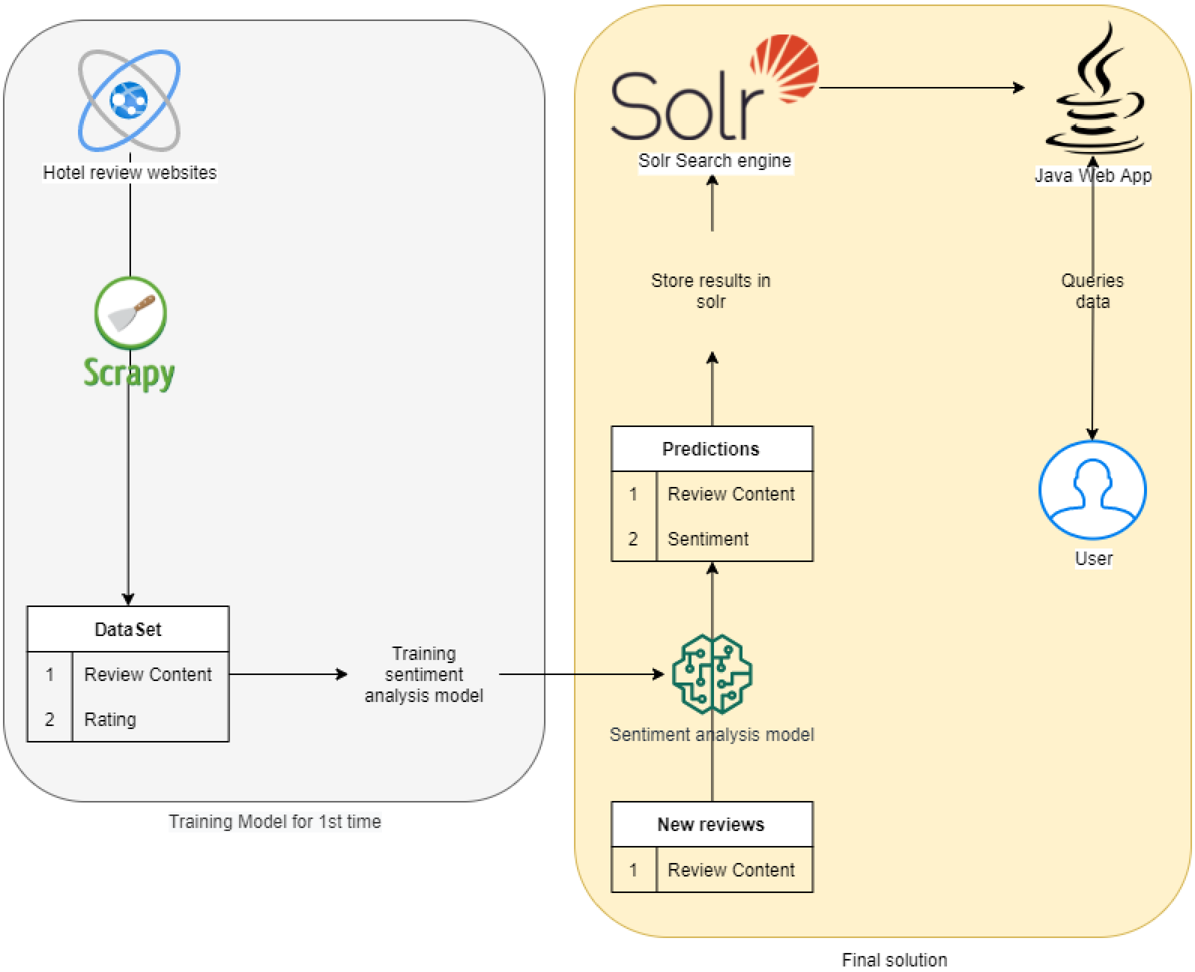

Figure 2: High level diagram of proposed model

This proposed system can be divided into five main subcategories. They are,

- Data gathering.
- Creating data sets.
- Data analysis.
- Storing Analysis data.
- Results visualization.

3.3. Data Set

Data (Hotel reviews) were collected from the booking.com web site. Since the booking.com is not providing an API to collect the reviews, Python data scrapping methods were used to collect the reviews. Python scrapy framework was used to achieve this. Total of 102899 Sri Lankan hotel reviews (English) were collected with the hotel name, review and review score.

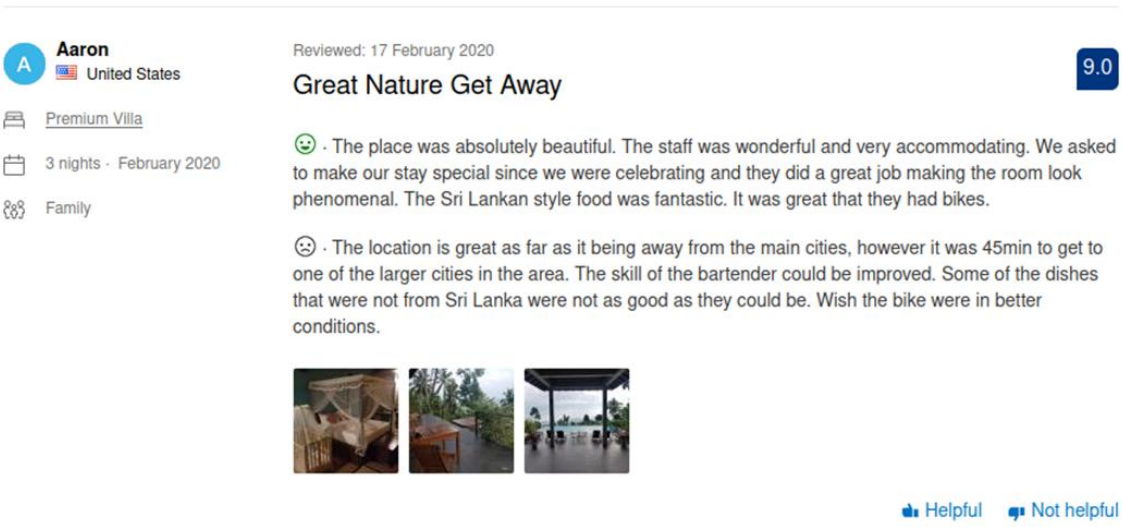

Figure 3: Sample review from Booking.com

```
hotel_name,rating,review_content,,,,,,
the-lion-pub-amp,10,"Closer to the main road, good on site restaurant, Cleanliness is great Vehicle noise in the road",,,,,,
the-lion-pub-amp,10,Good,,,,,,
the-lion-pub-amp,10,"Very convenient location, good spacious rooms and clean washroom",,,,,,
the-lion-pub-amp,10,"Location good, Good food, Nice Staff, Restaurant also good",,,,,,
the-lion-pub-amp,10,Its a very good hotel with comfortable and new furniture in the room. Staffs are attentive and helpful whe
beach.,,,,,,,
the-lion-pub-amp,10,Beautiful rooms with good amenities. Great to stay at this hotel due to its location. Overall a great plac
the-lion-pub-amp,10,"Beautiful rooms, Nice location. Great restaurant and pub on spot. I would recommend this place for others
the-lion-pub-amp,10,Excellent ,,,,,,
the-lion-pub-amp,4.6,Bed was good / comfortable. Had to wit for 20 min for them to do the check in. Thay did not know that the
not have water for a shower even. After I tolled them they offered an room change. First room I had to request  for towels anc
the-lion-pub-amp,3,,,,,,,,
the-lion-pub-amp,9,"Room was clean and nicely made.
Tea coffee facilities were available even at such a value for money rate.
 Hot water not available ",,,,,,
the-lion-pub-amp,7,Calm and good place,,,,,,
the-lion-pub-amp,9,"Spacious room with comfortable bed and bathroom,",,,,,,
the-lion-pub-amp,7,Staff was friendly. Allowed me to check out late.   You could hear the bus tooting and it echos in the roor
the lid up and the top of the flusher was dust. Floor was sticky. ,,,,,,
```

Figure 4: Collected reviews

3.4. Data Pre-processing

Before obtaining the features from the reviews text, the sentences were first preprocessed to remove any non-alphabetic characters from the sentences. All the characters were converted into lower case so that it is independent of the case being used in the texts. For each word it is then passed to the Porter Stemmer that removes any stop word from the sentences such as the, he, have, etc. it is required so that the model is not over fitted to these common words and instead looks to determine unique words that can help to determine its class value.

These 4 preprocessing techniques were used for the data preprocessing:

- Tokenizing the data
- Normalizing the data
- Noise removing

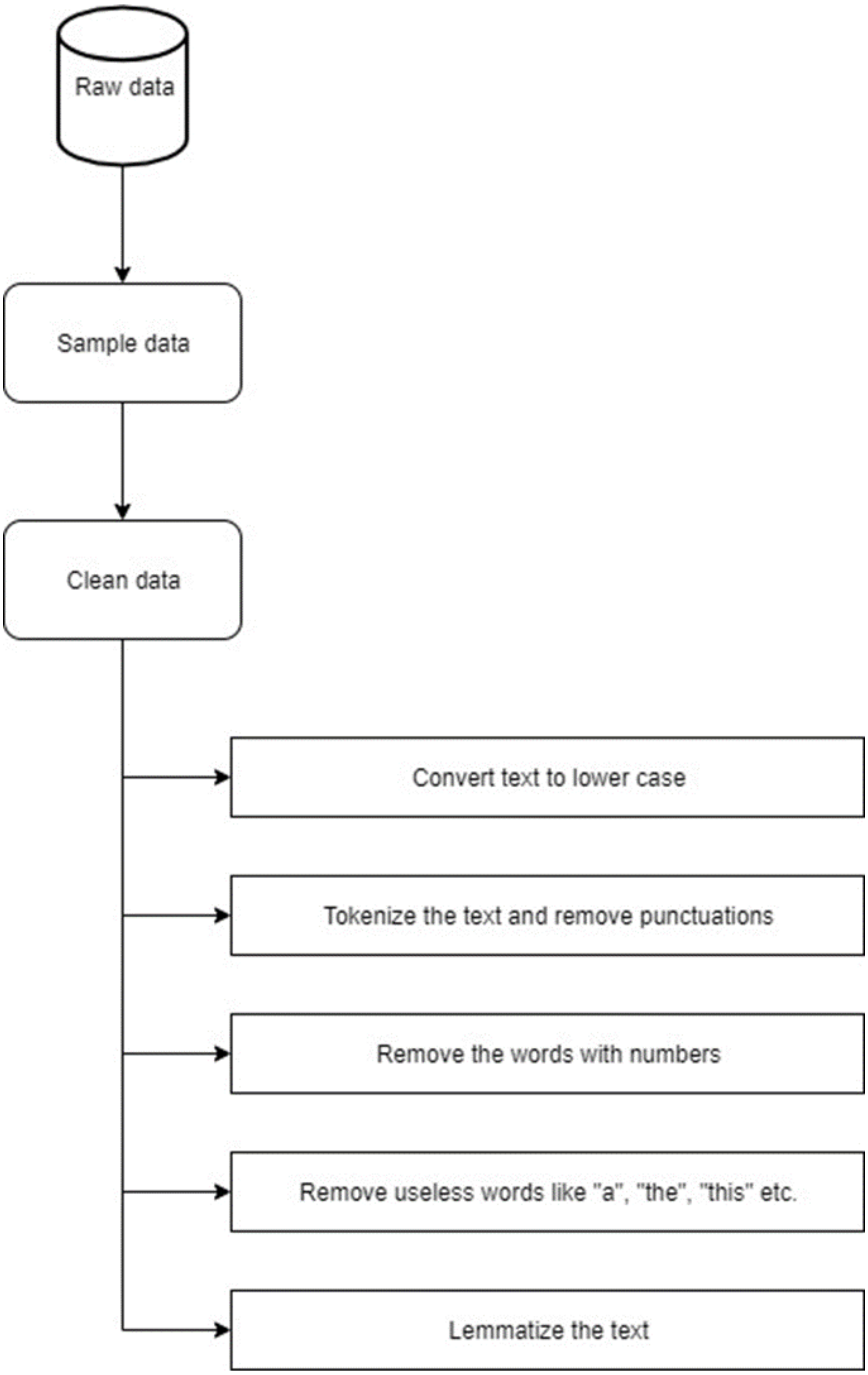

Figure 5: Data preprocessing

3.4.1. Tokenizing the Data

When the text is in its original form, machines cannot execute it. Therefore, it is essential to set the language to be easy for the machine to understand. Here we tokenize the data as the first part. An alphabetical order of text that serves as a single unit that it is called a token. Divide the strings into smaller sections called tokens. Here the text is divided using spaces and punctuation. Symbols can consist of hashtags, links, words, emotions, and individual characters.

3.4.2. Normalizing the data

The word "eat" has various forms, such as "eat", "eat" and "ate". In the normalizing, different forms of the word "eat" convert to the same form of the word "eat". The process of converting a word Its canonical form is what we call the normalization of the NLP. Normalization helps to combine group words with words Same meaning but in different forms. Will it be "eat", "ate" and "eaten" without normalization? Consider different words. Stemming and lemmatization, which are two techniques of normalization, were used in this work.

Stemming – This is a process of removing affixes from a word. Only works with simple verb forms.

Lemmatization – Normalizes a word with the context of the vocabulary and the morphological analysis of the words in the text.

Before the lemmatization, it is essential to determine the context of each word in the text. This can be realized using a tagging algorithm.

3.4.3. Noise Removing

Collected reviews data set contained garbage values and unwanted things for the classification. All of those things were removed. Stop words, hyperlinks, punctuations, special characters and emoji’s were removed.

- Empty reviews and reviews like nan, none, nothing and n a
- Reviews with only numbers
- Reviews with different languages
- All punctuations
- All the stop words
- Emoji’s

Figure 6: Noise removing

```python
def lemmatize_sentence(tokens):
    lemmatizer = WordNetLemmatizer()
    lemmatized_sentence = []
    for word, tag in pos_tag(tokens):
        if tag.startswith('NN'):
            pos = 'n'
        elif tag.startswith('VB'):
            pos = 'v'
        else:
            pos = 'a'
        lemmatized_sentence.append(lemmatizer.lemmatize(word, pos))
    return lemmatized_sentence
```

```python
def remove_noise(tweet_tokens, stop_words = ()):

    cleaned_tokens = []

    for token, tag in pos_tag(tweet_tokens):
        token = re.sub('http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+#]|[!*\(\),]|'\
                       '(?:%[0-9a-fA-F][0-9a-fA-F]))+','', token)
        token = re.sub("(@[A-Za-z0-9_]+)","", token)

        if tag.startswith("NN"):
            pos = 'n'
        elif tag.startswith('VB'):
            pos = 'v'
        else:
            pos = 'a'

        lemmatizer = WordNetLemmatizer()
        token = lemmatizer.lemmatize(token, pos)

        if len(token) > 0 and token not in string.punctuation and token.lower() not in stop_words:
            cleaned_tokens.append(token.lower())
    return cleaned_tokens
```

Figure 7: Noise removing from data

3.5. Sentiment Analysis Model

In this work, hotel reviews were classified into three sentiment types. They are positive, negative and neutral. Accuracy of the classification will be directly affected by the accuracy of the sentiment analysis model. So, it is important to use a model with highest accuracy to predict the polarity of the given review. There were three types of models trained in this work and finally used the model with highest accuracy. Naïve Bayes and convolution neural network were used to train these sentiment analysis models.

3.6. Naïve Bayes Classifier

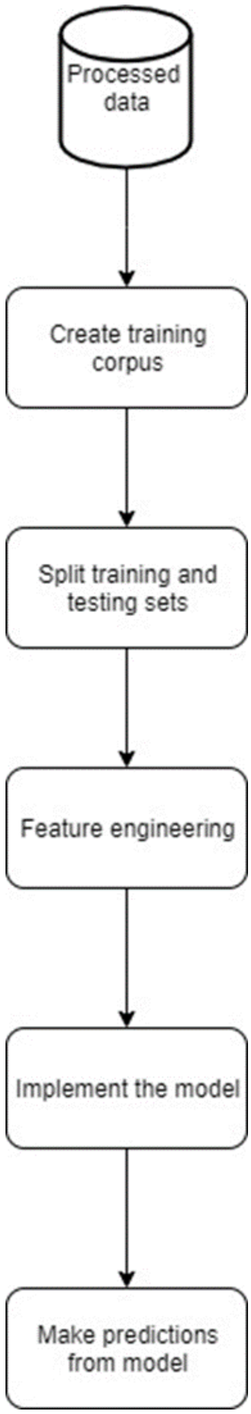

Figure 8: Sentiment Analysis Model

Naive Bayes is a very simple and fast algorithm for classifying a chunk of big data. For various applications such as spam filtering, text classification, emotional analysis, and recommendation programs, the Naive Bayes classification is used successfully. It uses the theorem of opportunities for Bayes in anonymous class predictions.

The Naive Bayes classification process is a simple and powerful task to differentiate in machine learning. The application of the Bayes theory with a strong sense of independence between the elements is the basis for the absurd division of the Bayes. When used for textual data analysis, such as Processing in Natural Language, the classification of Naive Bayes yields positive results.

Simple Bayes or independent Bayes models stand for other names of naive Bayes models. All these principles are based on the Bayes theorem. In practice, the Naive Bayes classifier uses the Bayes theorem. The power of the Bayes theorem is brought to machine learning at this stage.

P(A | B) – Posterior probability   P(A) - Class prior probability

P(B | A) – Likelihood   P(B) - Predictor prior probability

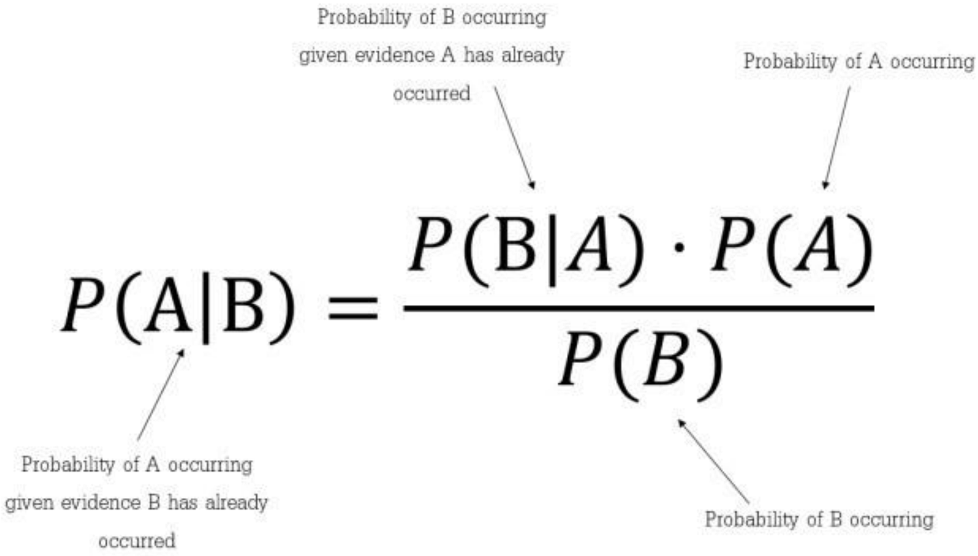

Figure 9: Review Classification

Total of 66000 hotel reviews were used to train the naïve bayes model. Set the train and test data according to the 7:3 ratio. But the final accuracy of this model was 0.69.

3.7. Neural Network Model

Neural networks learn in a similar way as brain learn, it takes series of input data and generates output which is then compared with the actual value to calculate error and based on the error value it then adjusts the weights of the connection by adding/subtracting value such that the desired output is generated.

Network starts with the small but not zero random values assigned to each connection weight in ANN. The dataset of features having same size as that of input layer is subjected to the network one by one, model produces some output and calculates error (difference between the output generated and the actual output in dataset). Network then attempts to minimize the error by adjusting the weights for each connection according to its contribution in the error value which is calculated by the process called "Back Propagation".

i.e. Error = Network produced output – Actual output from the dataset

Adjusted Weight = Weight + Error

$$\Delta \text{output} \approx \sum_j \frac{\partial \, \text{output}}{\partial w_j} \Delta w_j + \frac{\partial \, \text{output}}{\partial b} \Delta b$$

This Δoutput is calculated through backpropagation by applying gradient to the error generated with respect to the weight value so that it can generate how much the weight of the connection to be changed to obtain desired output from the network.

Network is kept on subjected to input of dataset and training goes on until an acceptance criterion is achieved, i.e.: when accuracy of the model is acceptable or the number of iterations for training has exceeded certain limit. Once the training is stopped, all the learning is stored in the form of weights of the connections which are then stored and used to produce output predictions.

Importance of the neural network over classical machine learning algorithms can be list down as below:

- Neural network is flexible to data as it can take missing, erroneous, and unrelated data and still can generate very good results.
- Layered architecture of neural network makes them able to be trained in parallel through other processing units such as GPU to speed up their training.
- Neural networks can learn amazing patterns in data and perform very well on its other related data as they are generalized.
- Fluctuations or Error in nodes will not affect the whole network as even in that case forward layers will be able to produce at least some data.
- Trained model does not need large space to store information as training is stored in the weight connections of the network.
- It supports transfer learning and weights can be stored in hard disk and reloaded in the

model and can learn further on new data without having to lose previously trained information. Example: Fine tuning.

- No mathematical equation needed for training as model learns by itself without taking care of its processing by mimicking human brain.

Network architecture of the neural network can be explained as below:

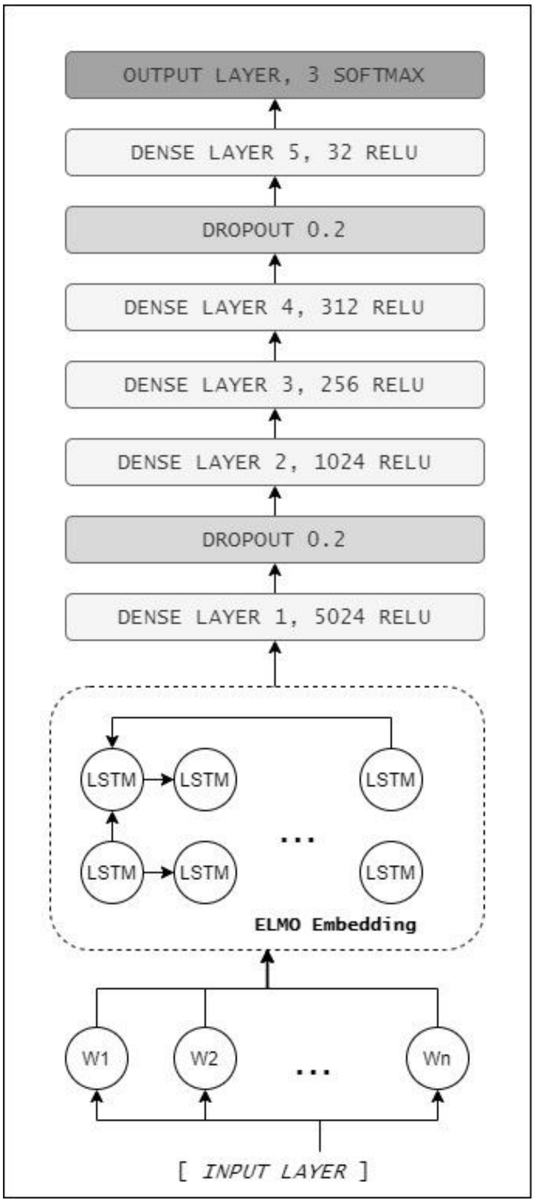

Figure 10: Network Architecture

Input Layer

Input layer contains 125 nodes equivalent to our feature vector size that receives input in its node as passes it to further network with weights being trained.

ELMO Embedded Layer

ELMO embedded layer which is a pretrained NLP layer provided by TensorFlow using bidirectional LSTM and produces contextual features of the tweets text. This layer produces fixed embeddings inside its each LSTM layer (3 are learnable), followed by mean-pooled feature for the input sentence.

Dense Layer 1

Large layer of 5024 nodes is connected directly to the ELMO embedded and with dropout 0.2 (to prevent overfitting through regularization). Activation function RELU is applied on this layer.

Dense Layer 2

1024 nodes layers were embedded with RELU Activation on the output received by Dense Middle Layer 1.

Dense Layer 3

256 nodes layers were embedded with RELU Activation on the output received by Dense Middle Layer 2.

Dense Layer 4

312 nodes sized layer is appended to the middle layers followed by 0.2 dropout to prevent overfitting due to dense sized network. Activation function of RELU is continued on this as well.

Dense Layer 5

Finally, a last middle layer of 32 nodes is applied having applied RELU activation function on each node’s output having input of the output produced by Dense Layer 4.

Output Layer

Output layer is a 3-node layer (as we have 3 classes) applied with SoftMax activation function that gives output in the form of probability belonging to either class between 0 to 1.

Output gets filtered that the value with less than 0.5 belongs to 0 class (which is fake in our labels) and 1 which represents real class. Training proceeds with categorical cross-entropy loss as there are two classes only and optimizes it with Adam.

Neural network model was also trained with 66000 reviews and 7:3 training: test data ratio. But the accuracy wasn’t increased significantly. The accuracy of the neural network model was 0.74.

3.8. Neural Network Model With Proposed Features

In order to classify hotel reviews from the data into the positive, negative and neutral, obtain textual features of text, visual marker features, sentiment related features, and emotional categories lexicon can be used as they contain significant information to identify text classes. Open-source libraries can be utilized in order to extract some of these features such as emotional lexicon, text, and sentiment features to identify some patterns in the data. Below, described the implementation details of the features.

3.8.1. Emotional Categories

This feature contains the list of emotional categories that are widely used and accepted by Word Emotion Association Lexicon. It extracts the frequencies of emotional categories from the given textual data. The significance of using this feature can be measured in terms of the concept that it is necessary to obtain emotions from the texts as they convey significant information to identify hatred speeches, over excited texts, encouraging comments, or such strong emotions, that are useful to classify them on the emotional basis. Following emotion categories were used for the implementation of this project:

1. Anger
2. Anticipation
3. Disgust
4. Fear
5. Joy
6. Sadness
7. Surprise
8. Trust
9. None

To obtain this **NRCLex** package has used to measure emotional frequencies from the text body that works by splitting words from the text and compares each word from the dictionaries of each emotional category to measure count of the features. It contains large dictionary of words for each emotion, i.e., around 27000 words based upon NLTK Wordnet data and National Research Council Canada (NRC) affect lexicon. For the feature data, this list of nine emotional categories having count of their respective emotion fetched on the given text has used.

3.8.2. Sentiment Features

To determine the sentiment measure of the text, the account should take into three important characteristics of the text, i.e., Polarity Value (PV), Sentiment Score (SS), and Semantic Similarity (SIM).

Polarity Value helps to measure the critical opinion for the input text as the negative value representing unfavorable evaluation on the text and positive being in the favor. Publicly available evaluation data can be utilized, (GitHub AFINN 2021) AFINN 11 for the text that has PV values given for each word and hence forming a dictionary with polarity scores for large set of English dictionary words. For each input text, we can split the words from the text and compare each word with AFINN words to obtain score and add it to the cumulative polarity score for the text and as a resultant, a final cumulative PV score for the text sentence will be taken as a feature.

Sentiment Score is used to obtain opinion decisions using the model developed by Hu Lui (Liu 2015). This works by splitting words and characters from the text and obtaining its character score from the pre-parsed model file and computing the resultant sentiment score for the complete text sentence.

Semantic Similarity is also taken into account for the sentiment analysis of the texts by using **WordNet** package. Synset (Synonyms Sets) from the WordNet module can be used to compute WuPalmer similarity value for the texts sentences. It works by obtaining synonyms list for each word of the text and measuring similarity with other of its words in the text sentences. This works for every word in the text and will compute cumulative similarity index value that is taken as a feature in this implementation model.

3.8.3. Visual Marker Features

These features are calculated statistically from the text to identify certain occurrences in the sentence and determine based on English grammar rules that might be helpful to get insights of nature of the sentences. Capitalization count, punctuation marks, emoticons, word length, exclamation marks, question marks, colons were taken as the character statistical measures and obtain verbs, nouns, adjectives, adverbs as the grammatical statistical features. Further, Hashtags and Mentions can also be taken into consideration for this feature.

Characters Statistics involve counting of capital characters, punctuation marks, emoticons, word length, exclamation, question marks, and colon, which are calculated by iterating through each character of the texts and incrementing their respective counts when occurrence is found.

Grammatical Statistics refers to obtaining statistical measures for the counting of some grammatical rules such as verbs, nouns, adjectives and adverbs. In order to implement this, making use of POS (Part of Speech) Tagger such as ARK Text NLP (Twitter Natural Language Processing) and its implementation in Python which is given in NLTK Pos Tagger that assigns part of speech tag categories to each of its word found in the grammar has used. For determining the verb, checking if the grammar contains **VB** tag, for nouns check if **NN**, for adjectives check if **JJ**, and for adverb check if **RB** is contained in the grammar tags has done. their instances were counted and append in the visual cues features to identify patterns of English grammar perspective from the texts.

3.8.4. Text Features

One of the important used features in this analysis will be the text feature as it extracts textual data from the sentences in the form of frequencies of the word. This will be done by using Count Vectorizer in python that tokenize the data and encode it in integers form so that the models can be trained. Keeping maximum features limit to be 100 so that it prevents the model to be over fitted on the textual data so that rather it makes use of the other sentiment and emotion lexical features as well. Count Vectorizer fits on the given dataset by selecting top 100 words with maximum frequencies and then calculates frequencies across all those words for each record and trains on them.

3.8.5. Model Training

After computing the feature vectors (Described in section 3.4), combine them all together to feed a final vector to input layer of our model architecture has done. NRC Lexicon gives 9 categories with emotions frequencies, Sentiment features produce 3 sized vectors with Polarity Value, Sentiment Score, and Word Similarity index, Visual marker feature is computed with 13 counts, and textual features is obtained with 100 vector size through Count Vectorizer. All are concatenated to produce resultant vector of 125 size.

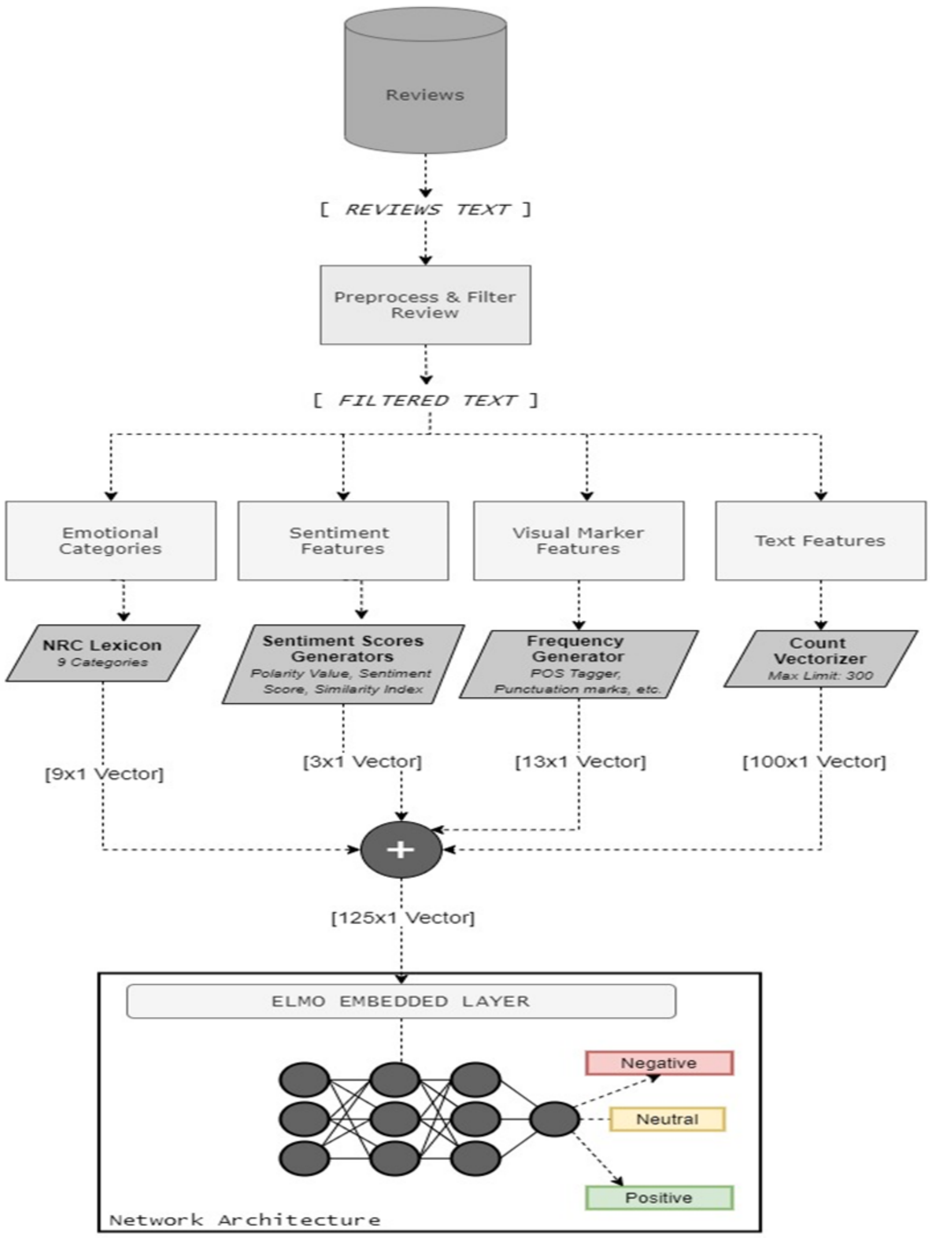

Figure 11: Model Training

Using the neural network model with the additional features, 0.89 of accuracy was achieved.

CHAPTER 4

EVALUATION AND RESULTS

In this project, mainly focused on to identify best machine learning model for sentiment analysis of Sri Lankan Hotel reviews. For this purpose, one machine learning algorithm and one neural network algorithm were used to generate models. These algorithms are Naïve Bayes and Convolution Neural Network algorithms. Initially Naïve Bayes model showed low accuracy with the hotel reviews and Convolution neural network model showed higher accuracy than the naïve model. So, the research was focused on to improve the accuracy of the CNN model. Finally, CNN model accuracy was improved after adding some additional features to it. Parameters like confusion matrix, accuracy, precision, recall and f1 score were calculated to compare the CNN model with improved CNN model.

```
Accuracy is: 0.6875
Most Informative Features
                 unclean = True           Negati : Positi =     72.9 : 1.0
                horrible = True           Negati : Positi =     51.3 : 1.0
                 disgust = True           Negati : Positi =     41.2 : 1.0
                   claim = True           Negati : Positi =     41.2 : 1.0
                    rude = True           Negati : Positi =     36.7 : 1.0
                  booked = True           Negati : Positi =     34.9 : 1.0
                    dirt = True           Negati : Positi =     34.9 : 1.0
                attitude = True           Negati : Positi =     32.3 : 1.0
             beautifully = True           Positi : Neutra =     29.7 : 1.0
                    unit = True           Negati : Positi =     28.5 : 1.0
```

Figure 12: Naive bayes results

```
Perc. Score 74.36%
F1 Score 76.48%
Recall 0.3271
Precision 0.74
Class  0 --ROC---> 0.51
Class  1 --ROC---> 0.51
Class  2 --ROC---> 0.49
Confusion Metric
[[    0    84  439]
 [    0   152  821]
 [    0 2064 9731]]
******************************************
```

Figure 13: CNN results without additional features

```
[14] Epoch 100/100
     1114/1114 - 4s - loss: 0.2497 - accuracy: 0.9037 - val_loss: 0.6822 - val_accuracy: 0.8908
     <keras.callbacks.History at 0x7fa12c1563d0>
```

```
[15] model_json = model.to_json()
     with open("cnn_model.json", "w") as json_file:
         json_file.write(model_json)
     model.save_weights("cnn_model.h5")
```

```
[16] model.load_weights('cnn_model.h5')
     print("## TESTING METRICS ##")
     loss, acc = model.evaluate(X_test, y_test, verbose=0)
     pred = model.predict(X_test)
     y_test = np.argmax(y_test, axis=-1)
     pred = np.argmax(pred, axis=-1)

     ## TESTING METRICS ##
```

```
[17] evaluate_model(y_test,pred)

     Perc. Score 89.15%
     F1 Score 86.18%
     Recall 0.4321
     Precision 0.89
     Class  0 --ROC---> 0.32
     Class  1 --ROC---> 0.45
     Class  2 --ROC---> 0.6
     Confusion Metric
     [[  123    80   320]
      [   52    71   850]
      [   41    99 11655]]
     ******************************************
```

Figure 14: CNN results with additional features

With the Naive Bayes model, only 0.68 accuracy could be achieved. The naive bayes model was given higher accuracy for the two-label text classification (Positive and Negative). That accuracy was 0.83. Since this sentiment analysis is doing using three classes, a new algorithm was needed to identify. After referring to some recent researches about sentiment analysis, decided to use a neural network algorithm for sentiment analysis in this research. Considering the usages and the internal implementation, decided to use the convolutional neural network (CNN) algorithm for this project.

```
Accuracy is: 0.8366666666666667
Most Informative Features
                     bit = True           Positi : Negati =     15.3 : 1.0
                 jacuzzi = True           Positi : Negati =     14.6 : 1.0
                   dirty = True           Negati : Positi =     12.8 : 1.0
                   never = True           Negati : Positi =     12.8 : 1.0
               fantastic = True           Positi : Negati =     12.5 : 1.0
                sigiriya = True           Positi : Negati =     11.6 : 1.0
                   order = True           Negati : Positi =     10.8 : 1.0
                    desk = True           Negati : Positi =      9.5 : 1.0
                  toilet = True           Negati : Positi =      9.5 : 1.0
                 perfect = True           Positi : Negati =      9.2 : 1.0
```

Figure 15:Accuracy of Naive Bayes model with two classes

With the first attempt at the classification with CNN, 0.74 accuracy could be achieved. That accuracy wasn't enough to go further with the project. So, needed to find a way to improve the accuracy of this neural network model. After studying more about the neural networks and sentiment analysis, came up with a solution to improve the accuracy of the current CNN model. It was to incorporate the text emotion extraction feature with the CNN model. With this solution finally, the sentiment analysis model achieved 0.89 of total accuracy. To compare further these two models, the below parameters were calculated to identify the most suitable sentiment analysis model for this project.

4.1. Confusion Matrix

The confusion matrix is a specific table layout also known as an error matrix. It allows to visualization of the performance of a machine learning algorithm.

Confusion matrix without additional features.

	Actual		
Predicted	0	84	439
	0	152	821
	0	2064	9731

Table 1: Confusion matrix without additional features

Confusion matrix with additional features.

	Actual		
Predicted	123	80	320
	52	71	850
	41	99	11655

Table 2: Confusion matrix with additional features

4.2. Accuracy

This is the most commonly used metric to judge a model and it doesn't provide a clear indicator about the performance.

Accuracy without additional features = 74.36 %

Accuracy with additional features = 89.15 %

4.3. Precision

Precision is the percentage of positive instances out of the total predicted positive instances. Precision value says, ‘how much the model is right when it says it is right’.

Precision without emotional features = 0.74

Precision without emotional features = 0.89

4.4. Recall/Sensitivity

Recall/sensitivity is the percentage of positive instances out of the total actual positive instances. This value tells, ‘how much extra right ones, the model missed when it showed the right ones.

Recall without additional features = 0.3271

Recall with additional features = 0.4321

4.5. F1 Score

F1 score is considered as the mean of precision and recall values. Higher F1 score would be better value.

F1 score = (2 * precision * recall) / (precision + recall)

F1 score without additional features = 76.48 %

F1 score with additional features = 86.18 %

After comparing all the parameters calculated above, best sentiment analysis model for Sri Lankan hotel reviews can be identified. When consider above calculated values, CNN model with embedded emotion extraction feature has the highest accuracy, precision, recall and f1 score. After embedding the emotion extraction feature, CNN model gives the highest performance as well as the highest accuracy.

CHAPTER 5

CONCLUSION AND FUTURE WORK

In 2019 tourism industry contributed 8.9 trillion US Dollars to the worlds gross domestic product. It is 10.3% from worldwide gross domestic product. 330 million of jobs that means 1 in 10 jobs around the world, directly and indirectly connected with the tourism industry. These data demonstrate the reputation of tourism industry around the globe. Why that much importance for the tourism industry is most of the people love travelling. People like to go and see beautiful, historical and ancient places around the world. They love to do adventures and they love to take experiences from different cultures around the world. In this journey travelers love to stay in best hotels and restaurants. How can they find the best hotel that fulfil their needs? From 15 to 20 years back people found best places by rumors. But now the situation is different. Most of the people find the best places by looking at the reviews in social media sites.

These days people can collect a lot of traveler appraisals in less than a blink of eye using internet facilities. People find it very difficult to find a suitable place to eat, visit and stay. Finding the right place, online reviews perform a big role that covers people's ideas, hobbies and desires. To get a general idea people should read all the other travelers' reviews completely but not randomly as there are positive and negative reviews. As a result of reading only a few reviews many people get the incorrect idea about the selected destination because it may lose a lot of the information that not read reviews contain, and they may encounter too many difficulties. Problems like these; to get a brief view of a hotel or restaurant visited by people, people should spend a lot of time reading and reviewing reviews of many places, studying all the reviews diminishes the yearning to travel of that explorer, and also analyzing the reviews to make the overall picture still lingers. As an appropriate remedy to this, a web system will be established to assist travelers to make opinions based on a rating system using Natural Language Processing technologies. From this web system, it will make easier for people to decide the most suitable hotels / restaurants for the trip. Apart from that it will decrease time people spend reading the reviews and it will make it easier to get a general idea about the quality of the destination and people will be able to choose the best place while decreasing the jeopardy of choosing a ruinous location as the destination.

Implemented system has capability to do the review classification and classified review summarization. System classifies the reviews as positive, negative and neutral. To classify the reviews both natural language processing and deep learning techniques were used. The CNN model was used to do the text classification. For improving the accuracy, emotions extractions have used. To train the model, a dataset which has created from this project, indicating Sri Lankan hotels data has used. Promised results were able to gain using the above model. Total accuracy of the model was 89%. The data set had lot of garbage values like empty fields, single words and reviews with different languages, before the usage, the data set was cleaned. Only the reviews written in English language was taken. When collecting the hotel reviews from the TripAdvisor web site, had to scrap the reviews from TripAdvisor. TripAdvisor does not provide an API to collect the reviews. As the future works of this project, willing to improve the text summarization by adding the abstractive text summarization. Ultimately willing to improve the system to auto generate some decisions to traveler.

REFERENCES

Dr. Ulrike Gretzel, Kyung Hym Yoo, Melanie purifoy, “Online Travel Review Study (Role & impact of Online Reviews) Laboratory for intelligent systems in tourism, Texas A&M University,” vol. 4.2, pp. 17–21, Feb. 2007.

S. Loh, F. Lorenzi, R. Salda˜na, D. Licthnow, A tourism recommender system based on collaboration and text analysis, Inf. Technol. Tour. 6 (3) (2003) pp. 157–165.

E. Bigdeli, Z. Bahmani, Comparing accuracy of cosine-based similarity and correlation-based similarity algorithms in tourism recommender systems, September, Management of Innovation and Technology, 2008. ICMIT 2008.4th IEEE International Conference on (2008) pp. 469–474.

M. Kenteris, D. Gavalas, A. Mpitziopoulos, A mobile tourism recommender system, June, Computers and Communications (ISCC), 2010 IEEE Symposium on (2010) pp. 840–845.

M. Al-hassan, H. Lu, J. Lu, Personalized e-government services: tourism recommender system framework, April, International Conference on Web Information Systems and Technologies (2010) pp. 173–187.

Y. Chu, H. Wang, L. Zheng, Z. Wang, K.L. Tan, TRSO: a tourism recommender system based on ontology, October, International Conference on Knowledge Science, Engineering and Management (2016) pp. 567–579.

C.S. Namahoot, N. Panawong, M. Brückner, A tourism recommendation system for Thailand using semantic web rule language and K-NN algorithm, Int. Inform. Inst. (Tokyo) Inform. 19 (7B) (2016) 3017.

X. Zheng, Y. Luo, Z. Xu, Q. Yu, L. Lu, Tourism destination recommender system for the cold start problem, KSII Trans. Internet Inform. Syst. 10 (7) (2016).

A.M. Kashevnik, A.V. Ponomarev, A.V. Smirnov, A multi model context-aware tourism recommendation service: approach and architecture, J. Comput. Syst.Sci. Int. 56 (2) (2017) pp. 245–258.

Q. Li, C. Wang, G. Geng, Improving personalized services in mobile commerce by a novel multicriteria rating approach, April, Proceedings of the 17th International Conference on World Wide Web (2008) pp. 1235–1236.

L. Liu, N. Mehandjiev, D.L. Xu, Multi-criteria service recommendation based on user criteria preferences, October, Proceedings of the Fifth ACM Conference on Recommender Systems (2011) pp. 77–84.

Mehrbakhsh Nilashia, Othman Ibrahima, Elaheh Yadegaridehkordib, Sarminah Samadc,Elnaz Akbari, Azar Alizadehfa, “Travelers decision making using online review in social network sites:A case on TripAdvisor”, Apr. 2018

Anania, Paolo (February 16, 2016). "Arbeiten bei Trivago: Keine Hierarchien und 15 Biersorten gratis". Lead Digital. Archived from the original on February 19, 2016.

Stanimara Yordanova, Dorina Kabakchieva, “Sentiment classification of Hotel Reviews in Social Media with Decision Tree Learning.” International Journal of Computer Applications (0975-8887), volume 158 – No 5, January 2017

Wanlian Tan (wanliang@stanford.edu), Xinyu Wang (xwang7@stanford.edu), Xinyu Xu (xinyu17@stanford.edu), “Sentiment Analysis for Amazon Reviews.”

Mita K. Dalal, Mukesh A. Zaveri, “Semisupervised learning based opinion summarization and classification for online product reviews.”

Stanley Loh, Fabiana Lorenzi, Ramiro Saldaña And Daniel Licthnow, “A Tourism Recommender System Based On Collaboration And Text Analysis”

Yohan Jo (yohan.jo@kaist.ac.kr), Alice Oh (alice.oh@kaist.ed), “Aspect and Sentiment Unification Model for Online Review Analysis”, Department of Computer Science KAIST Daejeon, Korea

Debnath Bhattacharyya1, Susmita Biswas2, Tai-hoon Kim1, “A Review on Natural Language Processing in Opinion Mining”, [1]Hannam University Daejeon, Korea. debnathb@gmail.com, taihoonn@empal.com [2]Computer Science and Engineering Department Heritage Institute of Technology Kolkata, India. bi.susmita@gmail.com

Dongjoo Lee[1], Ok-Ran Jeong[2] , Sang-goo Lee[1] “Opinion Mining of Customer Feedback Data on the Web”, [1] School of Computer Science and Engineering, Seoul National University Seoul 151-742, Republic of Korea (sglee@europa.snu.ac.kr), [2]Department of Computer Science, University of Illinois at Urbana-Champaign Urbana, IL, 61801, USA (orjeong@uiuc.edu)

Dhoha Almazro[1] (d_almaz@encs.concordia.ca), Ghadeer Shahatah[1] (g_shaha@encs.concordia.ca), Lamia Albdulkarim[1] (l_alabd@encs.concordia.ca), Mona Kherees[1] (m_khere@encs.concordia.ca), Romy Martinez[2] (romy.martinez@polymtl.ca), William Nzoukou[1] (w_nzouko@encs.concordia.ca), “A Survey Paper on Recommender Systems” [1]Concordia University, [2] Ecole Polytechnique, arXiv:1006.5278v4 [cs.IR] 12 (24) (2010)

Michael Kenteris, Damianos Gavalas, Aristides Mpitziopoulos, “A Mobile Tourism Recommender System”, Department of Cultural Technology and Communication University of the Aegean Mytilene, Greece, m.kenteris@ct.aegean.gr, {dgavalas, crmaris}@aegean.gr

Z. Bahramiana, R. Ali Abbaspoura,” AN ONTOLOGY-BASED TOURISM RECOMMENDER SYSTEM BASED ON SPREADING ACTIVATION MODEL”, School of Surveying and Spatial Information Engineering, College of Engineering, University of Tehran, North Kargar Ave., After Jalal Al Ahmad Crossing, Tehran, Iran (zbahramian, abaspour)@ut.ac.ir, November 2015

Dietmar Jannach, Zeynep Karakaya, Fatih Gedikli, “Accuracy Improvements for Multicriteria Recommender Systems”, TU Dortmund, Germany, DOI: 10.1145/2229012.222906, June 2012

Mehrbakhsh Nilashi[1], Dietmar Jannach[2], Othman bin Ibrahim[1], Norafida Ithnin[1] , “Clusteringand regression-based multi-criteria collaborative filtering with incremental updates”, [1]Faculty of Computing, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia, [2]TU Dortmund, Germany, 10 [26] Yoon Kim, “Convolutional Neural Networks for Sentence Classification”, New York University, yhk255@nyu.edu, 2014

www.cs.cmu.edu. (n.d.). Twitter Natural Language Processing -- Noah’s ARK. [online] Available at: http://www.cs.cmu.edu/~ark/TweetNLP/.

GitHub. (n.d.). sentiment_analysis/AFINN-111.txt at master · abromberg/sentiment_analysis. [online] Available at: https://github.com/abromberg/sentiment_analysis/blob/master/AFINN/AFINN-111.txt [Accessed 1 Aug. 2021].

Liu, B. (2015). Opinion Mining, Sentiment Analysis, Opinion Extraction. [online] Uic.edu. Available at: https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

APPENDICES

In this research project, first nearly 66000 of reviews were collected from the booking.com web site. Three sentiment analysis models were created using Naïve Bayes and CNN algorithms. Finally determined the most accurate model for the sentiment analysis. Google Collaboratory was used to train these models. Related materials to this project can be listed as below. This each material has been linked to the original resource:

- Dataset
- Naïve Bayes model
- CNN model without text emotion features
- CNN model with text emotion features