Analysis of Theme Impact in Consumer Reviews using Natural Language Processing Techniques

MSc Research Project
Data Analytics

Neha Suryawanshi
Student ID: x20169531

School of Computing
National College of Ireland

Supervisor: Majid Latifi

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Neha Suryawanshi
Student ID:	x20169531
Programme:	Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Majid Latifi
Submission Due Date:	17/12/2021
Project Title:	Analysis of Theme Impact in Consumer Reviews using Natural Language Processing Techniques
Word Count:	6389
Page Count:	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	31st January 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Analysis of Theme Impact in Consumer Reviews using Natural Language Processing Techniques

Neha Suryawanshi
x20169531

Abstract

The hotel industry is an economically growing asset for the United States of America (USA). It is a vast, diverse and an ever-expanding industry that provides people with services such as accommodation, food, leisure activities etc. What the guests think and post about the hotel has an impact on the potential consumers and thereby affecting the profitability of the hotels. Thus the guest reviews are crucial for hotels. For this very reason it becomes necessary to investigate the reviews, to identify any gaps in the business process. The reviews for hotels are available in a huge amount. Scanning manually through each and every review is impossible. NLP techniques such as topic modelling, text categorization, sentiment analysis etc., are efficient enough to process these large sets of textual data and produce a result comprehensible to a layman as well. This research focuses on identifying the aspects from the reviews so that the sentiment analysis can be done specific to the aspect.

1 Introduction

The rapid and vertical growth in the usage of world wide web has radically altered the shopping pattern of consumers. Dimensional Research carried out a survey and found out that 90% of the online consumers check the reviews from previous shoppers to learn more about the product or services (Gesenhues; 2013). The online reviews from other consumers help them make a better decision and hence it is now common to initially check the reviews before making the purchase.

1.1 Background

The hotel industry is a driving factor for the growth of the economy and consequently increasing job availability. They contribute to the local communities as 61% of the hotels i.e., more than 33,000 properties are regarded as small scale businesses. Over 250 local jobs are supported every night with 100 rooms occupied in the US. A study shows that guests on trips, spend $483 billion for their stay at hotels. (Longo; 2004). This motivates the research to study what the people like or dislike about their stay at the hotel particularly in the US using the aspect based sentiment analysis. It will help the industry to grow by making changes to the business process aligned to the interest of their guests.

1.2 Motivation

The success of hospitality industry is also partly dependent on the online reviews. A Subdivision of the hospitality industry is the hotel industry which provides people with services like lodging, food, spa, etc. Since booking of hotels is now available online, people tend to post reviews of their stay at the hotel stating their experiences and critically evaluating the hotel services.

In order to make the insight analysis simpler this research paper focuses on the aspect based sentiment analysis of reviews. Aspect based sentiment analysis is different from sentence based or document based analysis as it categorizes the review data by aspect or theme and further identifies the sentiment associated with it. The theme or aspect of the review is the component or attribute the guest is trying to evaluate which can be quality of rooms, taste of food, location of hotel and other facilities. This research will enable the guests and hoteliers to identify the theme of the review and its impact on the polarity of sentiment and aid their decision making for betterment of their services. It will further give hoteliers an idea of which of their service are up to the mark and help them identify gaps in their business process.

1.3 Research Question

To what extent NLP techniques can be used to identify the theme impact from the consumer reviews in order to provide a suitable business solution?

1.4 Research Objective

Following are the objectives of the research which will help answer the research question in an apt manner:

- **Objective 1 - Study state of art works:** Investigate and review the literature based on research related to sentiment analysis, aspect based sentiment analysis and restaurant rating prediction.
- **Objective 2 - Research methodology:** I propose a new research methodology to develop aspect based sentiment analysis using NLP techniques and python's NLP library like langdetect, TextBlob and Spacy.
- **Objective 3 - Implementation:**
  - Gather textual data and identify the aspect of the data using the statement context of the review.
  - Using NLP python library identify the sentiment associated with the aspect and generate a score.
- **Objective 4 - Evaluation:** Compare the results using the extrinsic metric - Coherence .

1.5 Contribution

The project's primary goal is to extract components of reviews for sentiment analysis. The majority of reviews are immediately analysed for sentiment analysis, which merely provides knowledge of whether the review is positive or unfavourable. However, using an aspect-based sentiment analysis improves the output's weight because it identifies which aspect of the review is favourable or bad. One of the project's unique approaches is to use Topic modelling to extract parts of the reviews. The project's key contribution is that, regardless of the domain, any business that collects consumer evaluations can use this implementation to analyse the data.

The remaining technical document is as follows: Chapter 2 discusses the related work of aspect based sentiment analysis with respect to online reviews. Chapter 3 elaborates the methodology of research. Chapter 4 includes the implementation of NLP for aspect based sentiment analysis as well as the ML methods for rating prediction. Chapter 5 states evaluation of NLP techniques used and results of machine learning algorithms. Finally Chapter 6 states the conclusion of conducted results which is followed by recommended future work.

2 Related Work

Research related to NLP, text analytics, sentiment analysis, aspect level sentiment analysis and ML methods are studied which are dominated by the topic of hospitality industry, its rating and customer experience. Further subsections have been documented as follows and include 2.1 - Sentiment Analysis, 2.2 - Aspect Level Sentiment Analysis and 2.3 - Hotel Rating Prediction.

2.1 Sentiment Analysis

The process of identifying whether the sentence or text has a positive, negative or neutral sentiment tied to it is known as Sentiment Analysis. In the field of text analysis, NLP is used commonly for sentiment analysis. NLP concepts can be understood even by a beginner programmer with the help of Natural Language Toolkit (NLTK) which is a python library for NLP. Žitnik et al. (2018) have used NLP to perform task such as text extraction, summarization and machine translation for information using NLP. Using the toolkit nut-IE of NLP was the main objective of their research. By using this toolkit they mainly carried out text language identification, sentiment analysis and language meaning extraction. This study was unable to provide any kind of comparison with other NLP toolkit but still qualifies for educational purpose. This research briefly explains how NLP can be put to use for language identification. The functioning of NLTK is elaborated by Lobur et al. (2011) on a deeper level. They mention that the vast library sources of NLTK are continuously updated over the time.

Rogers et al. (2020) proved in their research that the original transformation architecture can be used to carry out tasks such as text summarization, sentiment analysis and even build a question-answer system similar to a chat box. A Bidirectional Encoder Representations from Transformers (BERT) [1] model was proposed which first understood the language and its context. This model was further fine tuned. It could solve problems due to the training phase. They were able to show in their research that scope of original

[1] BERT: https://searchengineland.com/faq-all-about-the-bert-algorithm-in-google-search-324193

transformer architecture can be extended from only language translation to language context identification.

Machine learning is also one of the approaches used for textual sentiment analysis. Machine learning basically classifies the textual data by applying different algorithms such as Naïve Bayes, Support Vector Machine (SVM), Decision Trees etc. A hybrid model for the purpose of sentiment analysis performs better than that of individual machine learning algorithms and this very statement was proved true by was Mohamed Ali et al. (2019) as they carried out sentiment analysis using the machine learning approach on an IMDB dataset. This data set had half positive and half negative reviews. They successfully combined Long-Shot Term Memory (LSTM) with Convolutional Neural Network (CNN) to create the hybrid model CNN_LSTM and made comparisons of this hybrid model with CNN and LSTM separately. Post comparison of the experiment results, it was seen that the accuracy of the CNN_LSTM model was 89.2% which was better than that of the independent models i.e., CNN - 87.7% and LSTM - 86.74%. Experiments were also carried by Burns et al. (2011) on different datasets of different sizes and also on balance and unbalance datasets. Methods used for the experiments were Naïve Bayes and a dynamic language model which were further compared. They made use of a TV dataset and applied Bag of Words method to this dataset. The semantics of textual data was not very important for sentiment analysis of the text. This was proved by this research. It was also observed that both the methods under performed on balanced data rather than on the unbalanced data. On similar line Saito and Klyuev (2019) carried out a research using the Naïve Bayes classification model. This generative model was implemented to perform sentiment analysis for user reviews and the experiment also included other machine learning methods like SVM and K-nearest Neighbor (KNN). Using the mentioned machine learning methods the reviews were classified on review and sentence level and its polarity was identified which was either positive or negative. After the experiments were carried out the result suggested that Naïve Bayes method out performed SVM and KNN methods. The principle reason for Naïve Bayes to perform better was stated as the models navie assumption of independency. This also led to good results on large textual data. On the other hand, when Khan et al. (2019) carried out their research the results stated other wise. For their dataset SVM algorithm gave a higherg accuracy than that of Decision Tree and Naïve Bayes algorithm which was 90.3%. They proposed a framework which included gathering of data, its pre-processing as well as feature extraction and further carried out sentiment analysis using the mentioned machine learning algorithms.

Semantic Oriented Approach (SOA) is a dictionary-based approach which selects and identifies the sentiment words from the text. SOA contains dictionaries which are pre-developed and contain various phrases and words for which polarity is already assigned. This pre-developed dictionary helps it to identify the sentiment of the text (Wang et al.; 2014). Similar to SOA we have SentiWordNet and it is one of the resources which is very commonly used. Sentiment classification and opinion mining applications are supported by SentiWordNet as it is a lexical resource. Using SentiWordNet Singh et al. (2013) carried out review classification on a document level. Their method was designed to used the linguistic features containing adjectives, verb and adverb for the purpose of sentiment analysis. Using SentiWordNet, Agarwal et al. (2016) also performed sentiment analysis where news headlines was the text. They considered every part of speech in the news headlines for research purpose. A manually developed lexicon gave alike results to that of the SentiWordNet. This is proved by Ohana and Tierney (2009) in a study of text classification.

2.2 Aspect Level Sentiment Analysis

As seen in the above section application of sentiment analysis is widely studied in which the online content such as reviews, opinions and comments have been used as the textual data to analyse its sentiment towards the particular product or service. Similarly Muthukumarasamy (2014) and Hossain et al. (2018) have studied the reviews and opinions about the selected restaurant. Their main motivation was classify these reviews as either negative or positive as per the polarity using machine learning methods for sentiment analysis. They were able to identify only the polarity of the review and give an aggregated result. For more granulated results it is important to move beyond this crude method and employ a more efficient approach - Aspect/feature/theme based sentiment analysis.

The principle of this approach is to identify the aspect in the review and then further identify the sentiment associated with it. Aspect, feature or theme refer to the same concept and are used interchangeably. It is a particular property of the given product or service for which the consumers give their opinion or review about. This approach helps one to understand exactly what property or the product or service is liked or disliked by the consumer.

The given text's aspect can be extracted manually using different NLP techniques and is well portrayed by Akhtar et al. (2017) in their study. Use of techniques such as Parts-of-speech (POS) tagging, head word identification, word occurrence frequency, stop word removal, lemmatization etc., is done in order to identify the aspect of the textual data. This technique is comprehensive but very laborious and time consuming. It requires great amount of resources and time to get to a stage where one can begin the actual task of sentiment analysis. Using the technique of word embedding Mikolov et al. (2013) vectorized the list of identified aspect words. To do so they passed this list of words to a system - Word2Vec which found words that have semantically similar meanings and build a vectorized list of aspect words.

Post aspect identification the second step in this approach is to identify the polarity associated with it there by carrying out aspect-based sentiment analysis. Panchendrarajan et al. (2017) identified the aspects of a restaurant such as service, ambience, food, drinks etc., and further identified the sentiment associated to it. To generate a sentiment score they made use of the different lexical resources. These resources were employed on different context of the textual data. These contexts are comprised of document level, sentence level and N-gram level. The sentiment score helped identify the weak and strong qualities of the restaurant on a deeper level. This kind of output helps one understand the exact gaps in the business process and helps them bridge those gaps.

Besides the manual approach stated before we can use a pre-built dictionary for the purpose of sentiment analysis. A similar approach is followed by Peñalver-Martinez et al. (2014) when analyzing the viewer sentiments towards the entertainment industry. Making using of ontology - a set of entertainment related features or aspects such as actor name, movie name, genres etc., they carried out the study. This technique proves to be a powerful and efficient one and can be applied to other domain by just changing the ontology or dictionary to the related domain.

The above mentioned research and studies help understand that one can use various types of approaches or methods, either manual or with the help or pre-built dictionaries, for purpose of aspect identification like manual generation or using the dictionary or ontology related to the particular domain. In this research the dictionary or ontology based approach will be applied and evaluated. Discussion related to the same is in the following sections.

2.3 Hotel Rating Prediction

Comfortable hotel is a priority for tourist when planning a trip. To make the selection of a good hotel the first thing that one will check is the rating and review of the hotel. Based on previous experience of other people one can make a decision easily. Reviews found on online platforms are more authentic than those on the brochures of the hotels. Also if we compare ratings and reviews, reviews prove to be more reliable in describing the credibility of the hotel. As seen on the Tripadvisor website [2] for the hotel ibis Abu Dhabi Gate the rating is of 4.5 but one of the guest complained about the food quality while few complained about the check in and check out timings in their reviews. These reviews help give a better and detailed understanding of how the hotel is and helps the next guest to verify the credibility of the hotel.

Using the reviews and machine learning methods Liu (2020) predicted the rating by analysing the textual data. The textual data was handled using Tf-idf vectorization. Post the vectorization of textual data, machine learning methods like Naive Bayes (NB), Logistic Regression, Random Forest, and Linear SVM were utilised in order to predict the rating. The transformer-based models such as XLNet, BERT, RoBERTa and DistilBERT [3] were also used to predict which enhanced the performance by 6%. Issue of imbalance data was also resolved by Liu (2020) using the method of re-sampling. As per Luo and Xu (2019) the hidden semantic structure of the textual data is forsaken by most of the researchers. These hidden semantics include the emotions and linguistics that are delivered by the textual data. Food is the biggest contributor to the success of any hotel by Luo and Xu (2019) states that not only food but other services such as room quality, pool area etc., also contribute immensely to the success of the hotel. For the purpose of their research they considered mainly four aspects for the sentiment analysis viz., experience, food/taste, value and location. They also pointed out that people frown when they feel there is no value for the money. They made use of three hybrid machine learning methods which included NB, SVM and Fuzzy Domain Ontology (FDO).

To clean and pre-process the text methods such as case folding, punctuation removal, stopword removal, lemmatization and tokenization are generally used. Farisi et al. (2019) also used the similar methods and further applied the Multinomial Naïve Bayes model for sentiment analysis of the hotel reviews. They also made use of K-Fold Cross validation and Laplace smoothing along with Multinomial Naïve Bayes model. To analyse the sentiments of the reviews, the technique of Bag of words was also applied and they concluded that this technique was not feasible for a large dataset. They proved through their study that Tf-idf vectorization is an efficient technique for textual data analysis in large dataset. The performance and accuracy was also improved by using Laplace smoothing and K-Fold Cross validation.

[2] Tripadvisor website: https://www.tripadvisor.com/Hotel_Review-g294013-d2555801-Reviews-Ibis_Abu_Dhabi_Gate-Abu_Dhabi_Emirate_of_Abu_Dhabi.html

[3] Breif and comparison of XLNet, BERT, RoBERTa and DistilBERT methods: https://www.kdnuggets.com/2019/09/bert-roberta-distilbert-xlnet-one-use.html

2.4 Conclusion

After analysing various research papers and studies it is observed that the aspect based sentiment analysis and hotel rating predictions have always been carried out separately. In this research the study of aspect and prediction of rating based on the sentiment score of the aspect is carried out. For this purpose NLP techniques as well as machine learning methods are put to use.

The summary of related work is in the following Table 1:

Table 1: Summary of Related Work

Author	**Research Paper**	**Approach**	**Advantages**	**Drawbacks**
(Žitnik et al.; 2018)	NutIE - A modern open source natural language processing toolkit	Text extraction, summarization, machine translation, text language identification, sentiment analysis and language meaningextraction of textual data using the toolkit nut-IE.	Approach can be used for educational purpose	Does not demonstarte any comparison with other NLP toolkit
(Rogers et al.; 2020)	A primer in bertology: What we know about how bert works	An original transformation architecture - BERT is used which understands the text context and language to build a question and answer system resembling a chat box	Extend the scope of original transformation architecture from language translation to language context identification	None identified
(Saito and Klyuev; 2019)	Classifying User Reviews at Sentence and Review Levels Utilizing Naïve Bayes	Performed sentiment analysis of reviews on a sentence level using the machine learning methods SVM, KNN and Naïve Bayes classification.	Naïve Bayes out performed the two other methods and efficiently worked for large set of textual data.	Research proposed by Khan et al (2019) suggest otherwise. Their research suggests SVM out performs Naïve Bayes method.
(Mohamed Ali et al.; 2019)	Sentiment Analysis For Movies Review Dataset Using Deep Learning Models	Used a hybrid model by combining LSTM and CNN for sentiment analysis.	The combination of models gave a better result than that of the individual methods.	None identified
(Akhtar et al.; 2017)	Feature selection and ensemble construction: A two-step method for aspect based sentiment analysis	Used head word identification, POS tagging, stop word removal, lemmatization and word occurrence frequency for aspect identification and sentiment anaylsis of text.	The technique is comprehensive	The task is time consuming
(Panchendrarajan et al.; 2017)	Eatery - A multi-aspect restaurant rating system	Identified aspects like food, drinks etc., and then a sentiment score associated to it using a lexical resource. Analysis is of sentence level, document level and N-gram level	Helps identify the gaps in the business processes on a granular level.	None identified

3 Methodology

In implementing this research, a set of steps are followed. The process starts from selecting data, pre-processing it and further transforming it into usable format for the purpose of data mining. These steps are implemented in the research project and aim at generating knowledge related to aspects and their respective sentiments from the selected dataset. The flow of the applied methodology is seen in Figure 1 below. The rest of this section will explain the architecture of this study in brief overview:

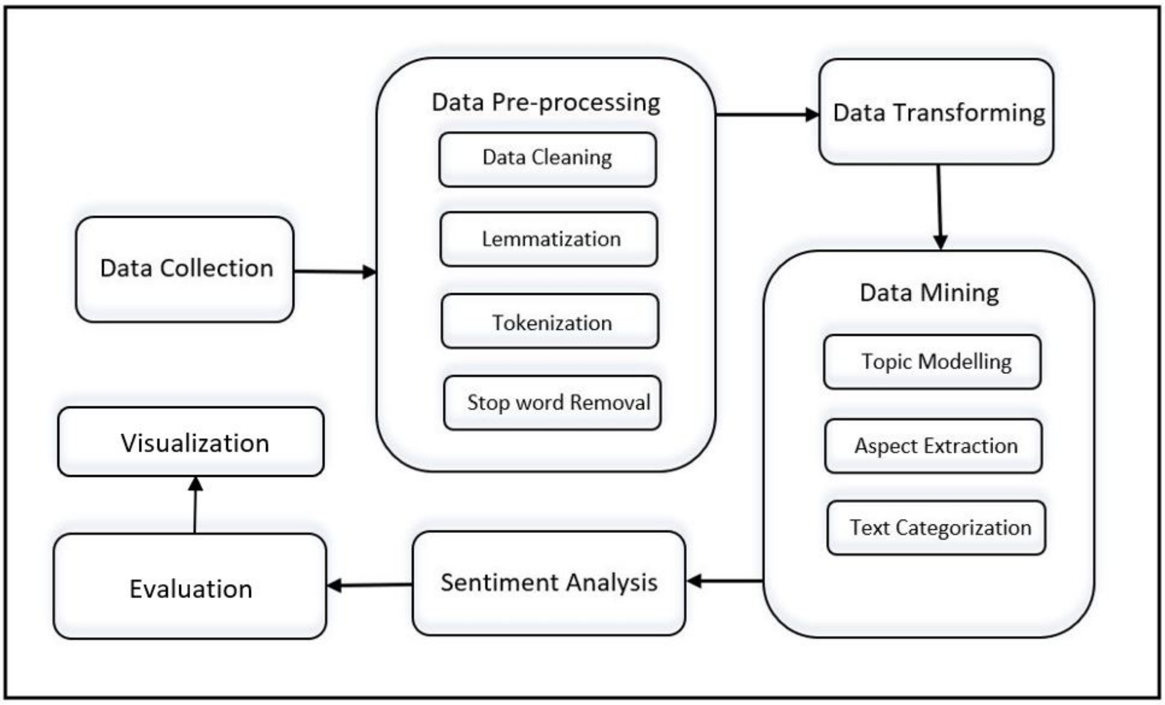


Figure 1: Methodology Flow

3.1 Data Selection/Collection

Numerous researchers have used hotel review data with the primary aim of sentiment analysis. In this research the hotel data from Datafiniti's Business Database is being used. This dataset is available on Kaggle [4] and contains data related to 1000 hotels primarily from the United States of America. This dataset is comprised of columns such as hotel name, category, location, rating, review, and more.

3.2 Data Pre-processing

Data is cleaned and pre-processed using the python library spaCy. The two columns 'reviews.text' and 'reviews.title' both have text describing the hotel hence they are combined into one column. Following pre-processing and cleaning operations are performed on the combined review column of the dataset:

[4] Hotel Reviews Dataset: `https://www.kaggle.com/datafiniti/hotel-reviews`

- **Handle NULL or missing values:** Delete or impute values based on the number of null/missing values.
- **Language detection:** Detect any non English review and remove them.
- **Case folding:** Convert all text to lower case.
- **Punctuation removal:** Remove all the punctuation marks except full stops (.)
- **Digit removal:** Remove any digits in the text.
- **Single character removal:** Remove any single characters as they do not help for aspect extraction or sentiment analysis.
- **Leading and trailing space removal:** Make the text more clean and usable by removing any leading or trailing spaces.
- **Lemmatization:** Group all words sharing the same meaning so as to analyse them as one single item. This single item is called as the word's lemma.
- **Tokenization of text:** In this step the text is converted into tokens, for example document or sentences into words. This helps to remove unnecessary tokens.
- **Stop word removal:** Any words which are not relevant or do not add any profound meaning to the sentence are removed.

3.3 Data Transformation

In this step of data transformation the clean text is converted into a dictionary. The conversion of reviews into dictionary made the task of text mining quiet simple. When performing text mining and sentiment analysis, dictionaries can be considered as maps which help to reach the destination or the goal in an efficient manner and in shorter time period.

3.4 Data Mining

Data mining is performed on the newly created dictionary. Topic modelling is performed on the reviews to identify the aspects and the further using text categorization the reviews are segregated as per the identified aspect.

3.4.1 Topic Modelling

To identify the review aspects topic modelling is performed on the created dictionary. This unsupervised method for text document classification is similar to the clustering performed on numeric data. The natural group of topics can be recognized using topic modelling. This method also is useful to find any hidden thematic structure of the text. Automatic searching, organizing, understanding and summarizing of massive files can be achieved using the method of topic modelling. It majorly helps with the following:

- Discovering the hidden themes in the collection.
- Classifying the documents into the discovered themes or aspects.

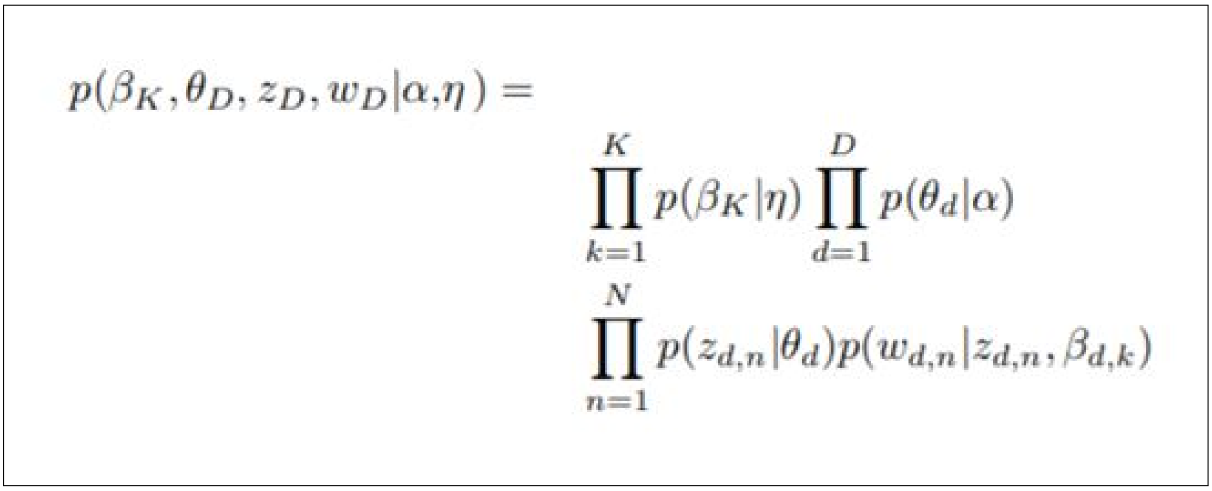
- Using the classification to organize/summarize/search the documents.

In 2003 Prof. David M. Blei developed a probabilistic topic modeling approach - Latent Dirichlet Allocation (LDA) which is also applied in the research for review aspect identification.

3.4.2 Aspect Extraction

The output of the LDA model are the various main topics of the dictionary provided. After running the enhanced LDA model the major topics are identified as the aspects of the reviews. Mainly these topics are the aspects which are criticised positively or negatively by the guest. The task of aspect extraction is simplified by the LDA topic modelling there by facilitating the text categorization activity. Using this technique it is possible to even identify the words related to the aspect. LDA follows generative probabilistic method which enables discovery of any hidden structure in the document. This generative process is as follows:

For every k topic in $\{1,...,k\}$, over the vocabulary V, β_k is a multinomial distribution which originates from Dirichlet distribution that is $\beta_k \sim$ Dir (η).
Representation of every document distribution over K topics also originates from the Dirichlet distribution θd $\sim$ Dir(α).
Where α denotes topic smoothing within the documents, and smoothing of words within topics is denoted by η. Following Figure 2 displays joint distribution of all hidden variables.



$$p(\beta_K, \theta_D, z_D, w_D | \alpha, \eta) = \prod_{k=1}^{K} p(\beta_K | \eta) \prod_{d=1}^{D} p(\theta_d | \alpha) \prod_{n=1}^{N} p(z_{d,n} | \theta_d) p(w_{d,n} | z_{d,n}, \beta_{d,k})$$

Figure 2: LDA

Where
β_K = topics
θ_D = proportions per-document
z_D = word topic assignment
w_D = words in document

3.4.3 Text Categorization

After aspects and its related words are identified, the similarity of each review with the newly identified aspects are calculated using spaCy library. This library is one of the fastest NLP libraries and is widely applied in today's industry. The similarity is calculated between the range of 0 to 1 where 0 denotes no similarity between the word vectors in the vector space while 1 denotes maximum similarity. For each aspect a .txt format file is created. The review with maximum similarity score with the aspect is stored in the respective aspect text file.

3.5 Aspect based Sentiment Analysis

Once the text files are ready for each aspect, the research proceeds with sentiment analysis on these files using TextBlob. This python library for Natural Language Processing (NLP) actively uses the Natural Language Toolkit (NLTK) to carry out tasks related to text analysis. Basically TextBlob provides a simpler interface to NLTK library as TextBlob is built upon NLTK. Many of the lexical resources are easily accessible due to NLTK library. It also facilitates users to work on tasks such as text classification. Complex analysis and operations on any kind of textual data can be easily supported by TextBlob and thus is used for sentiment analysis in this research.

There are two parts to TextBlob - Polarity and Subjectivity. The Subjectivity is a measure of the personal information or amount of personal opinion present in the textual data and this measure lies between [0,1]. The higher the value for subjectivity the higher is the amount of personal opinion rather than factual information. Polarity defines the negative or positive sentiment and ranges from [-1,1]. Where, -1 denotes negative while 1 denotes positive sentiment. For the purpose of this research only polarity of TextBlob is put to use as the aim of the research is to identify the aspect related sentiment. The semantic labels such as emojis or exclamation marks of TextBlob also enable the fine grained analysis of the textual data.

3.6 Data Evaluation and Visualization

The topic modelling done using LDA is evaluated using the coherence, a measure to evaluate the topics built by a topic model. For topic modelling, Rosner et al. (2014) mentions - Coherence is considered as an intrinsic evaluation method. If the sentences or facts support each other then they are said to be coherent. A set of facts is coherent when one can interpret that in a particular context a majority of the facts are contained (Mifrah; 2020). For example consider below sentences:

1. Trees need water to grow.
2. Many fruits grow on trees.
3. Trees are a source of oxygen.

In the above example all three sentences state facts about trees and hence these sentences are coherent. C_v coherence measure is efficient to attain the highest correlation among the provided data and hence adopted in the research to calculate the coherence.

The sentiment analysis carried out using TextBlob, on the respective aspect files are combined to identify polarity distribution for each aspect. This is also visualised using a bar chart and will be seen below in this technical report.

4 Implementation

This chapter describes the implementations carried out to fulfill the aim of the research. This chapter mainly covers details about how the steps mentioned in chapter 3 such as data pre-processing, data mining, aspect extraction, sentiment analysis etc., are performed.

4.1 Implementation of - Data pre-processing

The selected dataset has two columns - reviewtext and reviewtitle which describe the review from guests for the hotels. Since both the columns are useful for the analysis in the research they are combined into one column before any of the text cleaning steps are carried out. Also in the dataset the major of the reviews are observed to be in English language. Hence all the reviews in any other language than the English have been omitted from the research. To implement this step *langdetect* library is used which is a part of the Google's language-detection library. This particular library can support 55 languages and needs to be installed separately as it is not included in Python's standard utility modules. Further all the special characters, single characters, punctuation (except full stop), digits are eliminated using the replace function and the processed text is saved in *'review.txt'* file. The frequent words in the newly created file can be seen in the word cloud Figure 3 below:

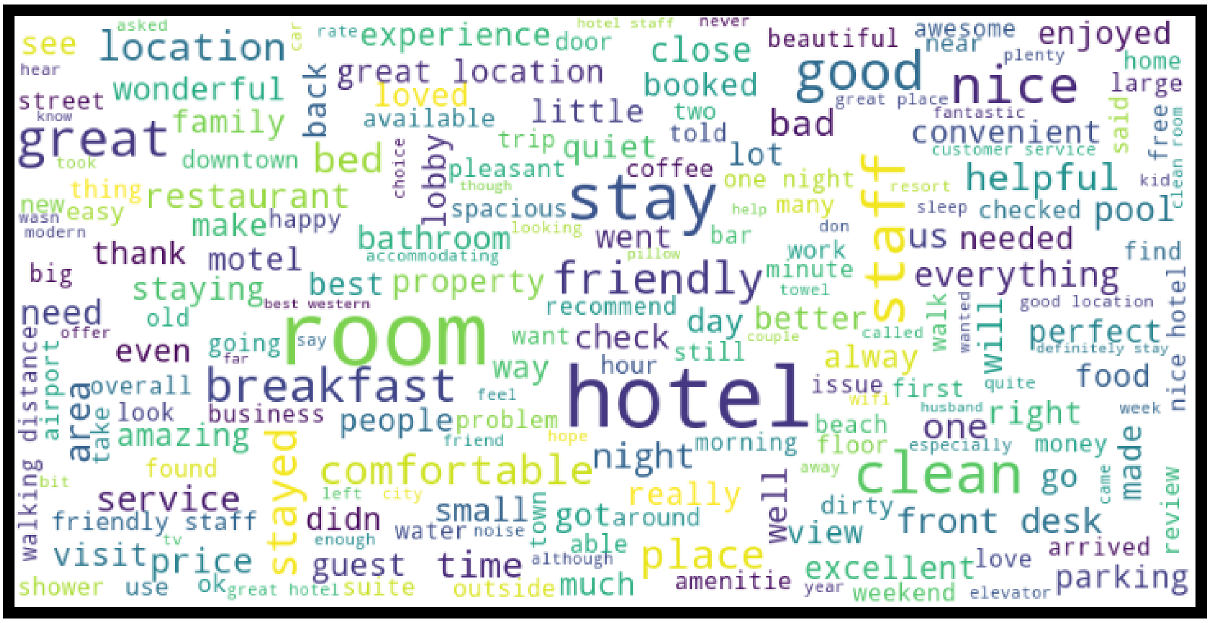


Figure 3: Word Cloud for Reviews

As seen in the above Figure 3 the words such as hotel, room, stay, great breakfast, location, nice etc., occur frequently in the reviews. But this word cloud is not efficient enough to identify the aspects of the reviews and hence topic modelling approach has been applied in the research which is described at a later stage in this technical report.

The clean text is then lemmatized and tokenized and any stop words present in the text are eliminated. This step is performed using the spaCy library. Lemmatization is

the process of breaking down words into their base forms which is known as 'lemma'. Example: the words talk, talking, talks indicate the same activity but have different spelling structure. To avoid any confusion in our algorithm such words are lemmatized under a single lemma. The text is then tokenized to chop it down into pieces called tokens and further any of the stop words which do no contribute in text analysis are eliminated. Along with the default list of stop words of spaCy library, few more unusable words were discovered during topic modelling. These extra unusable words shown below are added to the stop words present in spaCy library in order to remove them from the textual data and obtain a clean set of text. The custom stop words are:
'try', 'use', 'don', 'hair', 'child', 'th ', 'st', 'con', 'look', 'check', 'time', 'love', 'I', 'rancho', 'come', 'motel', 'day', 'find', 've', 'moregreat', 'weekend', 'find', 'family', 'business', 'stop', 'want', 'trip', 'kid', 'need', 'lot', 'sit', 'close', 'option', 'ok', 'air', 'cool', 'morenice', 'thank', 'return', 'take', 'new', 'entire', 'husband', 'choose', 'daughter', 'inn', 'hotel', 'minute', 'right', 'care', 'way', 'include', 'go', 'ask', 'pm', 'get', 'hour', 'leave', 'expect', 'definitely', 'hard', 'like', 'feel', 'tell', 'ask', 'change', 'didn', 'moregood'.

The text after lemmatization, tokenization, stopword removal, special character and digit removal is saved in file *goodtext.csv* file for further use. As mentioned before the full stops were not removed and are used as a separator while appending text in the *goodtext.csv* file. This enables to keep the clean text as per sentences in the review and makes the aspect wise file creation easier. Using this newly created file, dictionary is created which is then used for topic modelling. Any full stops present in the dictionary are removed as they do not contribute in the task of topic modelling.

4.2 Implementation of - Topic Modelling and Aspect Extraction

In order to identify the aspects of the reviews topic modelling is carried out. LDA's generative probabilistic method enables the discovery of even the hidden topics/aspects and hence is applied in this research for aspect extraction. The gensim LDA model uses the created dictionary and corpus. The dictionary is required to build the bag for words which contributed to the creation of corpus. Randomly the first LDA model is run with number of topics as 4. This gave a clear picture of the topics/aspects of the reviews but still it needed more enhancements. These enhancements are achieved using the extrinsic evaluation metric of coherence. The value of coherence lies between 0.1 to 0.9 where model with 0.1 coherence is inadequate and 0.9 value is an over fitting value.

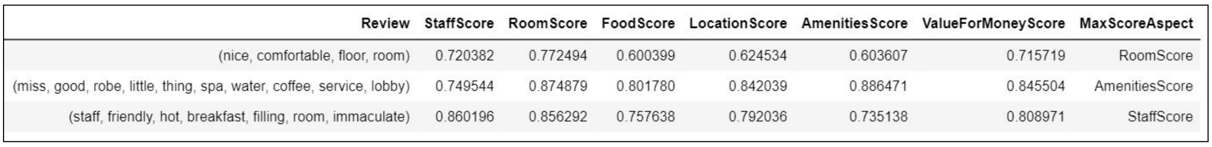
The optimal number of topics for LDA model was found out to be 10 with alpha = 0.01 and eta = 0.9. This model helped discover two more aspects - amenities and value for money. After analysing the topics from the LDA model the following aspects and their related frequently occurring words were extracted. These aspects and related words are in the following Table 2:

Table 2: Extracted Aspects and their Related Words

staff	**room**	**food**	**location**	**amenities**	**valueForMoney**
staff	room	food	outside	amenities	spend
friendly	clean	water	place	pool	price
helpful	comfortable	breakfast	town	pet	free
desk	quite	coffee	area	lobby	offer
welcome	bathroom	restaurant	travel	western	service
employee	bed	eat	location	parking	rate
attentive	shower	complimentary	far	microwave	extra
professional	door	egg	walk	fridge	value
rude	smell	cookie	street	table	pay
accommodating	small	continental	road	book	deal
checkin	large		distance	view	reasonable
super	hot		drive	shuttle	charge
owner	old		mile	elevator	budget
execellent	carpet		away	kitchen	refund
courteous	towel		convenient	smoke	
	spacious		locate	cigarette	
	suite		airport	cook	
	maintain		access	maintenance	
	accommodation		beach	internet	
	sleep		downtown		
	light		point		
	stain		metro		
	upgrade		state		
	pillow		station		
	bedroom		harbor		
	sheet		shopping		
	sink		mall		
	comfy				

4.3 Implementation of - Text Categorization

After the successful identification of six aspects, six text files are created, each for its respective aspect. These text files are populated with the reviews that have similarity with the aspect. The similarity between reviews and aspects have been calculated using the similarity method of the spaCy library. For each aspect a list of its similar words are created and matched with each clean review and a similarity score is calculated. The review is classified under the aspect with maximum score.This enables each sentence to be added into a file only once and avoids any overlaps. Following Figure 4 is an example from the textual data where the score is assigned based on the similarity with the aspect and the review is then classified under the aspect with maximum score. In the first row the review is about room and hence the room aspect has the highest similarity score of 0.772494. Due to this the review is classified as review under the room aspect. Similar is for the rest of the two row as well. They describe the amenities and staff and have the highest similarity score for amenities and staff respectively hence are categorized accordingly.



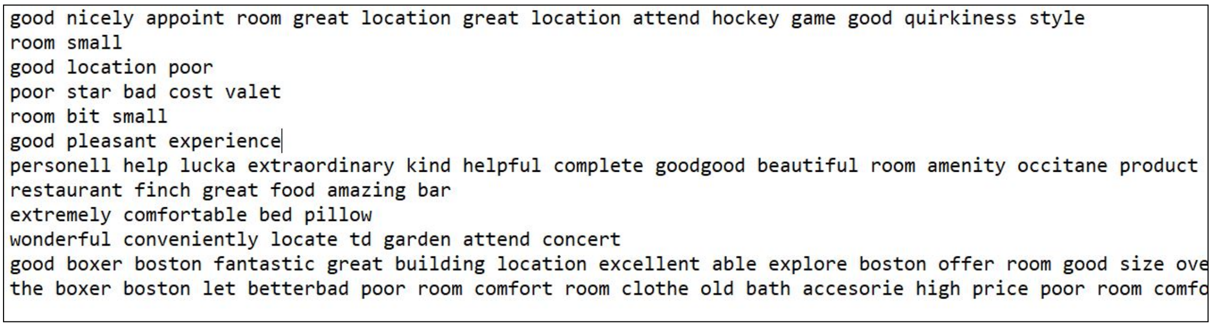
Review	StaffScore	RoomScore	FoodScore	LocationScore	AmenitiesScore	ValueForMoneyScore	MaxScoreAspect
(nice, comfortable, floor, room)	0.720382	0.772494	0.600399	0.624534	0.603607	0.715719	RoomScore
(miss, good, robe, little, thing, spa, water, coffee, service, lobby)	0.749544	0.874879	0.801780	0.842039	0.886471	0.845504	AmenitiesScore
(staff, friendly, hot, breakfast, filling, room, immaculate)	0.860196	0.856292	0.757638	0.792036	0.735138	0.808971	StaffScore

Figure 4: Review Aspect Score

The text files that are created for each of the aspect are as follows:

reviews_for_staff.txt	reviews_for_room.txt	reviews_for_food.txt
reviews_for_location.txt	reviews_for_amenities.txt	reviews_for_valueformoney.txt

Following is Figure 5, a snapshot of the file reviews_for_room.txt



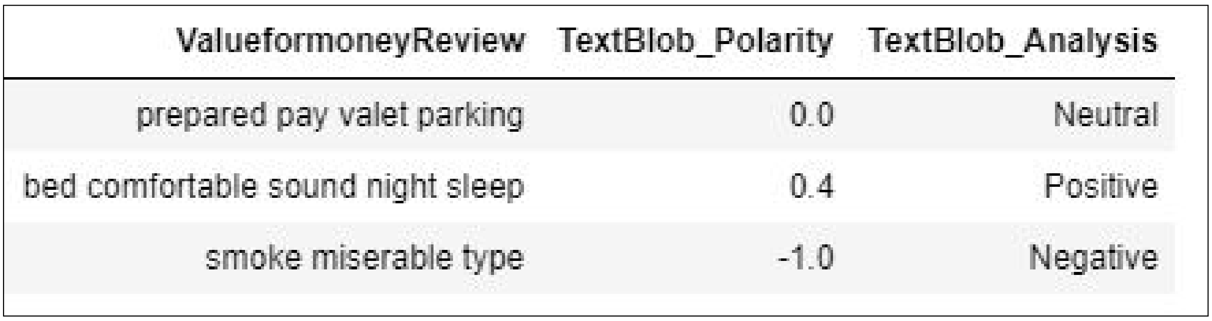
```
good nicely appoint room great location great location attend hockey game good quirkiness style
room small
good location poor
poor star bad cost valet
room bit small
good pleasant experience
personell help lucka extraordinary kind helpful complete goodgood beautiful room amenity occitane product
restaurant finch great food amazing bar
extremely comfortable bed pillow
wonderful conveniently locate td garden attend concert
good boxer boston fantastic great building location excellent able explore boston offer room good size ove
the boxer boston let betterbad poor room comfort room clothe old bath accesorie high price poor room comfo
```

Figure 5: Text File for Room Aspect

4.4 Implementation of - Sentiment Analysis

The research is continued by performing sentiment analysis on each of the aspect file. TextBlob is a python library which is a part of the NLTK used for sentiment analysis. As mentioned before the research has implemented only the polarity of TextBlob to identify if the review is positive negative or neutral. The polarity score lies between [-1,1] and reviews with polarity score less than 0 are categorized as negative, reviews with polarity score 0 are categorized as neutral and reviews with polarity score more than 0 are categorized as positive. Following Figure 6 illustrates how TextBlob has assigned the sentiment polarity score for 'Value for Money' aspect.

A bar chart for the same aspect as seen in Figure 7 displays 800 reviews that have a neutral sentiment, 700 plus reviews with positive sentiment and 200 plus reviews with negative sentiment.



ValueformoneyReview	TextBlob_Polarity	TextBlob_Analysis
prepared pay valet parking	0.0	Neutral
bed comfortable sound night sleep	0.4	Positive
smoke miserable type	-1.0	Negative

Figure 6: Review Polarity

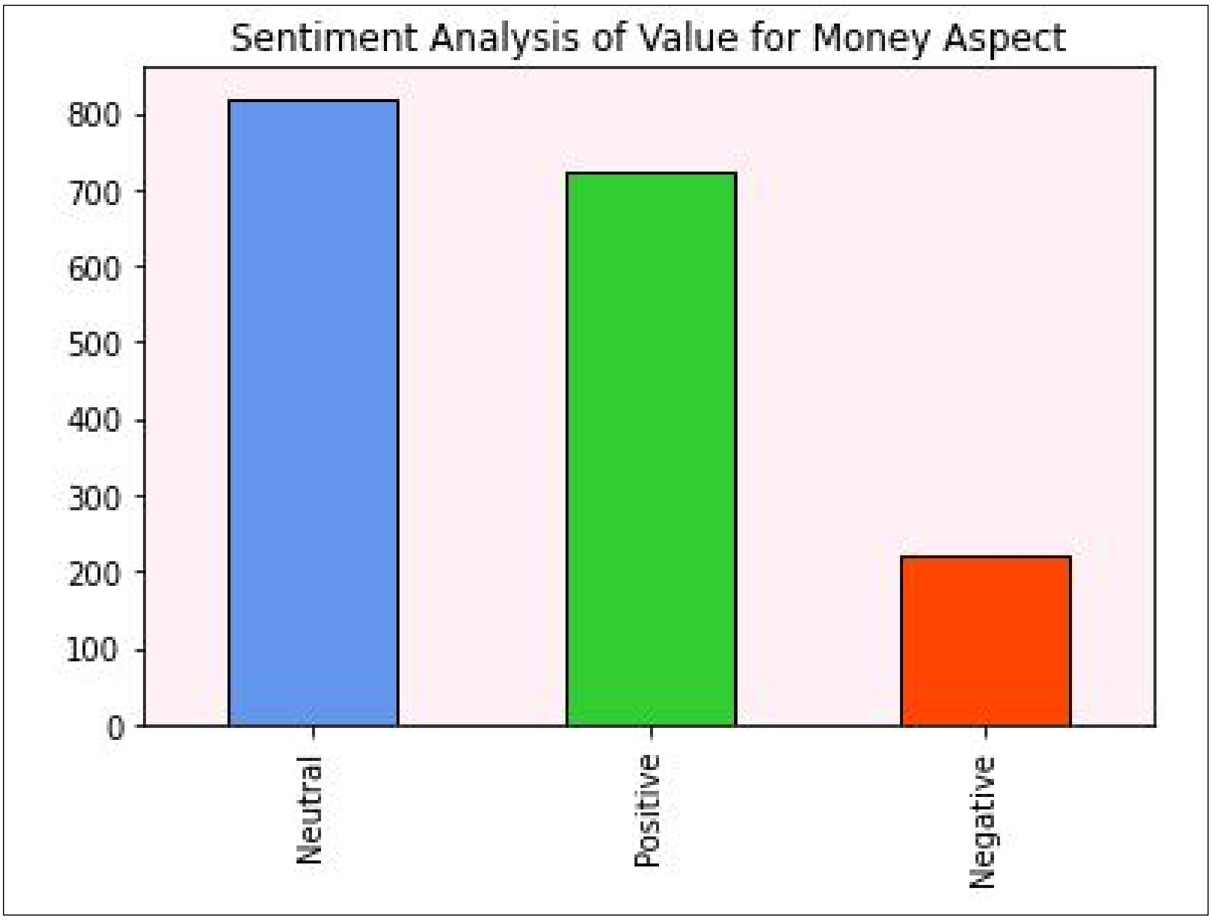
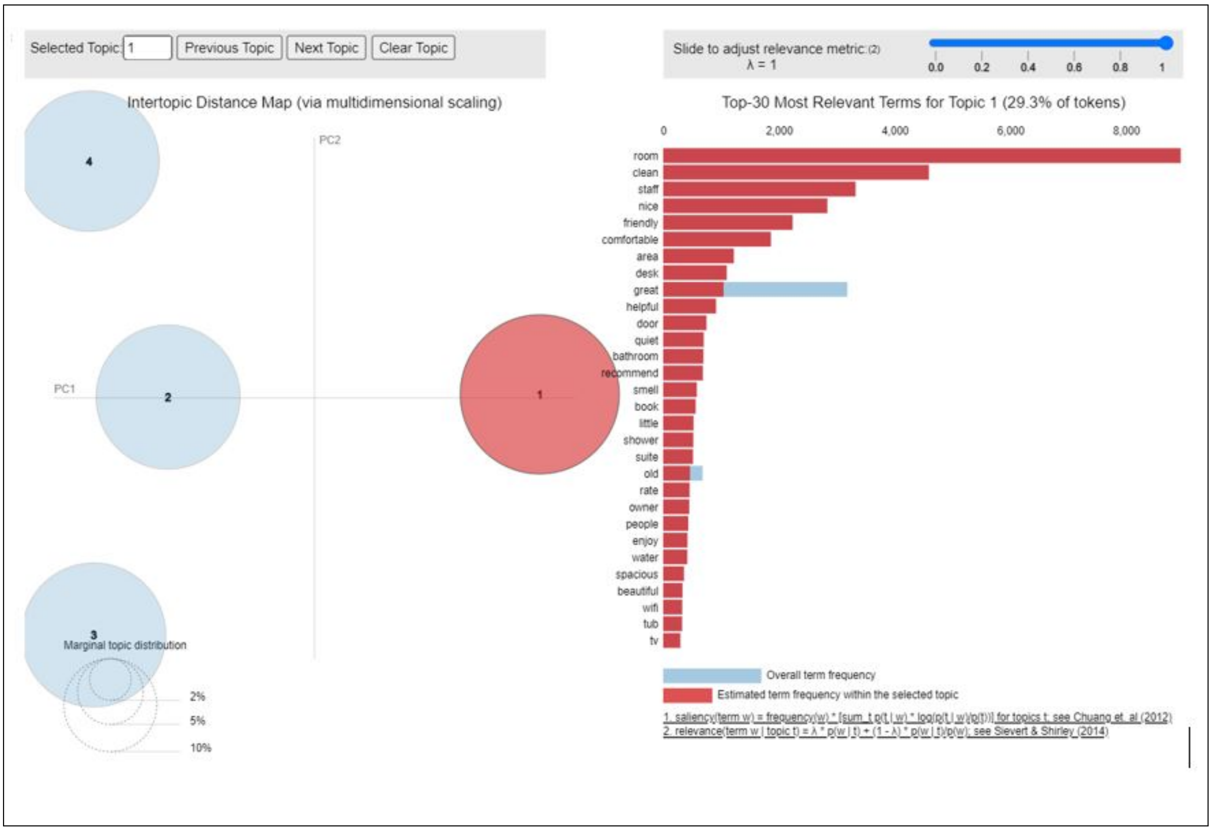


Figure 7: Sentiment Analysis for Value for Money Aspect

5 Evaluation

5.1 Baseline LDA Model

The LDA model for topic modelling is first run with parameters number of topics as 4. The hyperparameters alpha and eta kept default 1.0/num_topics for this baseline model. The chunksize parameter is responsible for controlling the number of documents processed at a time. Increasing its value accelerates the training process but the chuck of document should fit into the memory. For the baseline model chucksize is set to 100. The passes parameter is set to 10 and controls the number of times the model is trained on the corpus. This passes word can also be referred to as epochs and is necessary to set this number to a higher value. The baseline model majorly showed 4 aspects or topics - Room, Staff, Location and Food. The topics or aspects are also visualized using the pyLDAVis and are seen in Figure 8. The extrinsic evaluation metric - coherence for this model is 0.2983 which is extremely poor. Also identification of the topics is complex in this model.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 8: LDA model 1

5.2 Hyperparameter Tuning

The hyperparameters for LDA model are

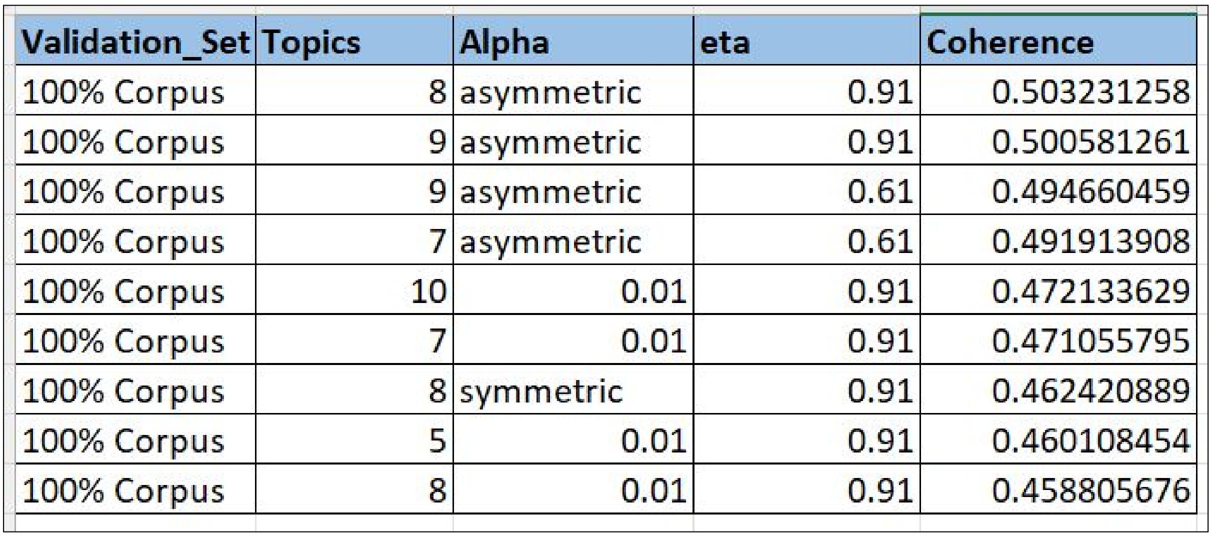
- k - Number of topics
- alpha - A Dirichlet hyperparameter which is Document-Topic Density

- eta - A Dirichlet hyperparameter which is Word-Topic Density

Taking one parameter at a time the tests are performed in sequence. The other parameters are kept constant. The tests are run on two differed corpus sets. And the choice of metric for performance comparison is c_v. This metric is basically about these four parts:

1. Data segmentation into pairs of words.
2. Probability of word pair.
3. Confirmation measure calculation which decides how strongly a set of words is related to another set of words.
4. Calculate overall coherence which is the aggregation of the confirmation measure.

The optimum model is identified by iterating over a range of topics, alpha and eta values. The result is saved in a file lda_tuning_results.csv. This file is analysed by sorting the values in descending order of the coherence value. The top ten observation with highest coherence is displayed in the following Figure 9. It is seen that the optimal number of topics is 10 while the value for alpha is 0.01 and eta is 0.9.



Validation_Set	Topics	Alpha	eta	Coherence
100% Corpus	8	asymmetric	0.91	0.503231258
100% Corpus	9	asymmetric	0.91	0.500581261
100% Corpus	9	asymmetric	0.61	0.494660459
100% Corpus	7	asymmetric	0.61	0.491913908
100% Corpus	10	0.01	0.91	0.472133629
100% Corpus	7	0.01	0.91	0.471055795
100% Corpus	8	symmetric	0.91	0.462420889
100% Corpus	5	0.01	0.91	0.460108454
100% Corpus	8	0.01	0.91	0.458805676

Figure 9: Coherence Values Based on Aspect

Using the tuned hyperparameters the LDA model is run again. This model helped discover the hidden two topics - Amenities and Value for money. The topic modelling using LDA enabled to identify appropriate aspects for the purpose of sentiment analysis with the coherence value of 0.4721 while the baseline model's coherence value was 0.2983. The hyperparameter tuning yielded a significant improvement of 62%.

5.3 Aggregation and Visualization of Sentiment Analysis

The sentiment analysis for each file is combined to show at one glance how each of the aspects are rated by the guests. The collated results are shown with the help of a bar chart as seen below in Figure 10. The maximum number of positive reviews are for the rooms while they are least for the value for money aspect. But this is true only in the big picture. If we consider the number of reviews available for the value for money aspect the number of positive and neutral reviews are similar and number of negative reviews are less. The guests are very impressed by the hotel staff and they hardly have a negative sentiment towards this aspect. For the food as aspect the most of the guests are neither impressed nor unhappy. If we look at the bar chart at a glance it is clear that the hotels in the US are doing well with respect to aspects like staff, rooms and location but need to up their game with respect to aspects like food, amenities and value for money as the number of positive reviews are less comparatively.

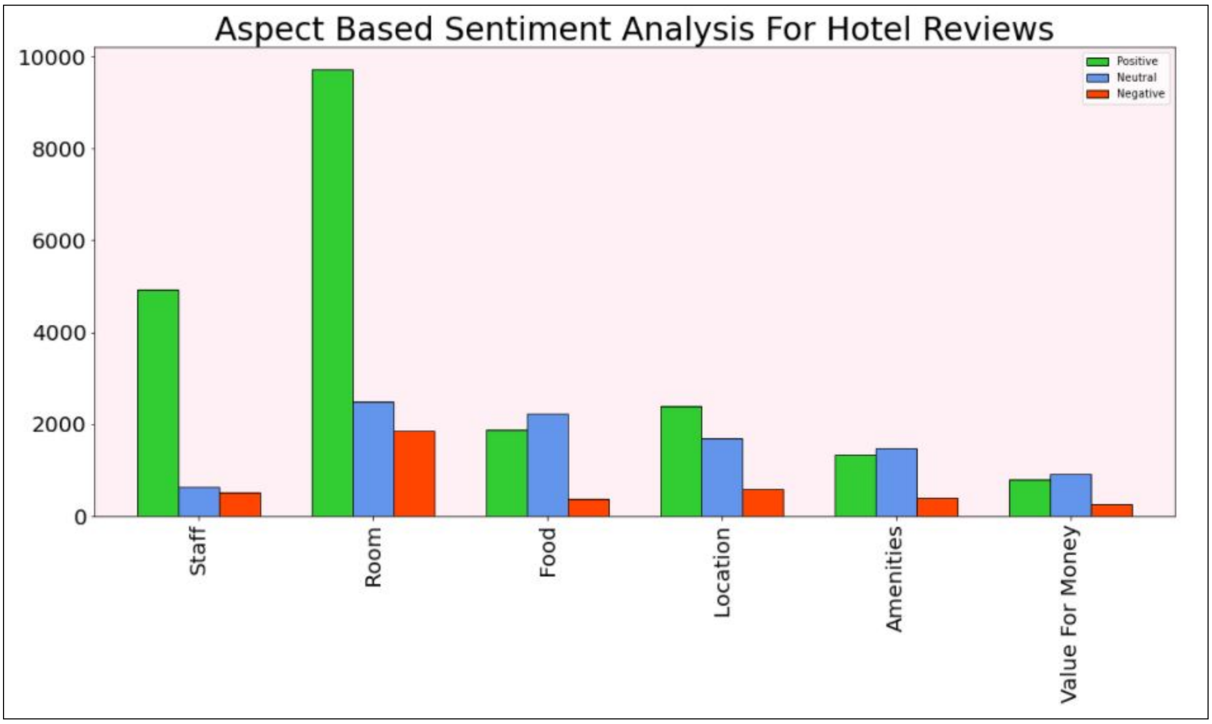


Figure 10: Sentiment Analysis

5.4 Discussion

The research methodology is unlike the methods proposed by the previous researchers. In this research the use of topic modelling made it possible to extract the aspect prior to its sentiment analysis. Also the use of similarity method from spaCy library for text classification is also one of the novel approach applied in this research. This research is successfully able to produce an implementation which can be used by any domain which makes use of user reviews. E-commerce or movie reviews are few of the many domains which can make use of this implementation to understand the customer perspective towards the products or services offered and take further action based on the analysis. This enables such domains to learn from the results and make further adjustments to their business to increase their profit margins and also improve customer satisfaction. The consumers also can benefit from this application to get a brief idea about the product or service before proceeding to the purchase.

6 Conclusion and Future Work

The analysis of the theme impact on the hotel reviews has been accomplished using various NLP techniques such as topic modelling, text categorization and sentiment analysis. This task made use of python libraries like spaCy and langdetect for text cleaning and pre-processing. The LDA topic modelling facilitated the identification of aspects of the textual data (user reviews for hotel). Hyperparameter tuning improved the performance of the LDA model by 62% and also helped discover two of the hidden aspect - amenities and value for money. Also the similarity method from spaCy library aided the categorization of reviews into each newly created text files as per the identified aspect during topic modelling. The sentiment analysis of each aspect file using TextBlob python library led to the learning of the impact each aspect has on the consumers. It also revealed how each aspect has an impact on the hotel guests which again will help the business for further plan of action.

For future work the text categorization task will be implemented using the TfidfVectorizer and compare its output with this research. Also as mentioned before this implementation will be applied to a dataset of different domains which make use of reviews for consumer consideration. Implementation of POS tagging will also be done to verify any performance improvement in aspect identification.

7 Acknowledgement

I would to express my gratitude to my supervisor Mr.Majid Latifi who helped me through my research project. His continuous support and guidance throughout the final semester motivated me and helped me to have a good learning experience. I would also like to thank my parents and friends who have constantly supported me and have been my morale boost.

References

Agarwal, A., Sharma, V., Sikka, G. and Dhir, R. (2016). Opinion mining of news headlines using SentiWordNet, *2016 Symposium on Colossal Data Analysis and Networking, CDAN 2016*, Institute of Electrical and Electronics Engineers Inc.

Akhtar, M. S., Gupta, D., Ekbal, A. and Bhattacharyya, P. (2017). Feature selection and ensemble construction: A two-step method for aspect based sentiment analysis, *Knowledge-Based Systems* **125**: 116–135.

Burns, N., Bi, Y., Wang, H. and Anderson, T. (2011). Sentiment Analysis of Customer Reviews: Balanced versus Unbalanced Datasets, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 6881 LNAI, Springer, Berlin, Heidelberg, pp. 161–170.

Farisi, A. A., Sibaroni, Y. and Faraby, S. A. (2019). Sentiment analysis on hotel reviews using Multinomial Naïve Bayes classifier, *Journal of Physics: Conference Series*, Vol. 1192, IOP Publishing, p. 012024.

Gesenhues, A. (2013). Survey: 90% Of Customers Say Buying Decisions Are Influenced By Online Reviews.

Hossain, F. M., Hossain, M. I. and Nawshin, S. (2018). Machine learning based class level prediction of restaurant reviews, *5th IEEE Region 10 Humanitarian Technology Conference 2017, R10-HTC 2017* **2018-January**: 420–423.

Khan, D. M., Rao, T. A. and Shahzad, F. (2019). The Classification of Customers' Sentiment using Data Mining Approaches, *Global Social Sciences Review* **IV**(IV): 146–156.

Liu, Z. (2020). Yelp Review Rating Prediction: Machine Learning and Deep Learning Models.

Lobur, M., Romanyuk, A. and Romanyshyn, M. (2011). Using NLTK for educational and scientific purposes, *2011 11th International Conference - The Experience of Designing and Application of CAD Systems in Microelectronics, CADSM 2011*, pp. 426–428.

Longo, K. (2004). American Hotel & Lodging Association: All Together Powerful 33,000 + PROPERTIES ARE SMALL BUSINESSES $245B SUPPORTING $1.1T OF U.S. SALES Includes hotel revenue, guest spending and taxes.

Luo, Y. and Xu, X. (2019). Predicting the helpfulness of online restaurant reviews using different machine learning algorithms: A case study of yelp, *Sustainability (Switzerland)* **11**(19).

Mifrah, S. (2020). Topic Modeling Coherence: A Comparative Study between LDA and NMF Models using COVID'19 Corpus, *International Journal of Advanced Trends in Computer Science and Engineering* **9**(4): 5756–5761.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013). Efficient estimation of word representations in vector space, *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, International Conference on Learning Representations, ICLR.

Mohamed Ali, N., El Hamid, M. M. A. and Youssif, A. (2019). SENTIMENT ANALYSIS FOR MOVIES REVIEWS DATASET USING DEEP LEARNING MODELS, *International Journal of Data Mining & Knowledge Management Process* **09**(03): 19–27.

Muthukumarasamy, G. (2014). Sentiment Analysis of Restaurant Reviews Using Bagged Ensemble Classifiers, *. . . Journal of Artificial Intelligence and . . .* (1): 17–23.

Ohana, B. and Tierney, B. (2009). Sentiment classification of reviews using SentiWordNet, *9th. IT & T Conference* .

Panchendrarajan, R., Ahamed, N., Sivakumar, P., Murugaiah, B., Ranathunga, S. and Pemasiri, A. (2017). Eatery - A multi-aspect restaurant rating system, *HT 2017 - Proceedings of the 28th ACM Conference on Hypertext and Social Media*, Association for Computing Machinery, Inc, pp. 225–234.

Peñalver-Martinez, I., Garcia-Sanchez, F., Valencia-Garcia, R., Rodríguez-García, M. Á., Moreno, V., Fraga, A. and Sánchez-Cervantes, J. L. (2014). Feature-based opinion mining through ontologies, *Expert Systems with Applications* **41**(13): 5995–6008.

Rogers, A., Kovaleva, O. and Rumshisky, A. (2020). A primer in bertology: What we know about how bert works, *Transactions of the Association for Computational Linguistics* **8**: 842–866.

Rosner, F., Hinneburg, A., Röder, M., Nettling, M. and Both, A. (2014). Evaluating topic coherence measures.

Saito, Y. and Klyuev, V. (2019). Classifying User Reviews at Sentence and Review Levels Utilizing Naïve Bayes, *International Conference on Advanced Communication Technology, ICACT*, Vol. 2019-Febru, Institute of Electrical and Electronics Engineers Inc., pp. 681–685.

Singh, V. K., Piryani, R., Uddin, A. and Waila, P. (2013). Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification, *Proceedings - 2013 IEEE International Multi Conference on Automation, Computing, Control, Communication and Compressed Sensing, iMac4s 2013*, pp. 712–717.

Wang, H., Yin, P., Zheng, L. and Liu, J. N. (2014). Sentiment classification of online reviews: Using sentence-based language model.

Žitnik, S., Drašković, D., Nikolić, B. and Bajec, M. (2018). NutIE - A modern open source natural language processing toolkit, *2017 25th Telecommunications Forum, TELFOR 2017 - Proceedings* **2017-January**: 1–4.