Feature Comparison of Yelp Restaurants Based on Sentiment Analysis

W.N.N De Silva

169307b

Faculty of Information Technology

University of Moratuwa

September 2020

Feature Comparison of Yelp Restaurants Based on Sentiment Analysis

W.N.N De Silva

169307b

Dissertation submitted to the Faculty of Information Technology, University of Moratuwa, Sri Lanka for the partial fulfillment of the requirements of the Degree of Master of Science in Information Technology.

September 2020

Declaration

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

Name of Student   Signature of Student

Date:

Supervised by

Name of Supervisor   Signature of Supervisor

Date:

Acknowledgement

Foremost, I would like to express my sincere gratitude to my advisor Senior Lecturer, Mr. Saminda Premaratne for the continuous support of my MSc study and research, for his motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis.

Last but not the least, I would like to thank my family, my parents for supporting me spiritually throughout my life and to all of my university colleagues who helped me in this venture.

Abstract

Today the internet has given a huge space for people to explore new experiences. It has made lives more comfortable. Global information can be retrieved within seconds over any topic. Advance of such internet technologies made people to share own experiences over various topics. The extensive use of social media, forums, blogs or various e-commerce platforms have made this easier. People give opinions on various products or services based on their experiences. These reviews affect marketing strategies of online businesses. Also, comments are important for the staff or the owners as well. Positivity or the negativity of these comments is important. But thousands of reviews are collected per a day. In order to get an idea about the reviews one has to spend many hours reading them. This is an impossible task. Hence there should be an easier way to explore huge number of reviews with in seconds. This research is based on implementing a system to facilitate this task. The domain of the business is restaurants. People often search for good restaurants. Most of them are used to explore various customer reviews to find the best one. The solution will be a hybrid sentiment-based system which aids customers to find good restaurants in a particular city. Customer review datasets of various restaurants in a particular city are collected from Yelp.com. The reviews are analyzed based on popular restaurant features such as price, quality of food, ambience or service. The main functionality of the system is to deliver a sentimental comparison between various restaurants over their features. Several NLP tasks with machine learning techniques such as multilabel Naïve Bayes model, SVM model, deep learning convolution neural networks and word embeddings are used in this research.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction to Systems based on Natural Language Processing Techniques

1.1 Introduction

1.1.1 Background

NLP based technologies play a great role when dealing with natural languages. Today the internet is full of text reviews, ratings chat conversations, emails, messages etc. The advance of internet technologies has taken people sharing own ideas and opinions in to a higher level. Customers are continuously making comments on products and services which affects marketing campaigns of online businesses. Billions of reviews are collected per a day. Reading these comments one by one is a time-consuming task. NLP related technologies have gained a huge recognition in performing various tasks on these customer reviews. There are various types of online business domains. Restaurants are considered as a popular one. Many people search for restaurants using internet. Also, they try to search for opinions of others based on experiences to find a suitable one. Hence customer review mining has become a popular research topic. Most of the review mining tasks are based on NLP technologies. Customer opinions reveal sentiments towards a particular topic or a subject. Each review contains information on a particular aspect or more. The sentiments towards an opinion can be positive or negative. Positive reviews are good for a business. Also, there can be fake or miscellaneous reviews. Valid reviews should be collected for a specific task. Task of finding sentiments of reviews is known as sentiment analysis or opinion mining. Considerable number of researches has been implemented based on sentiment analysis as well. Advanced applications such as translators such as Google translator, personnel assistants such as apple Siri, grammatical checkers in word processing applications are great inventions implemented using natural language processing technologies. Human languages are not easier to understand for a computer. A language contains many features such as syntaxes, semantics, discourse, speech, sentiments etc. Hence NLP is considered more complex when identifying those features separately using the computer.

Restaurants search and making reservations are currently a popular online business. Restaurant reviews are even more popular when searching for a good one. Hence Review mining with restaurants give many opportunities for researchers in academic and industry as well. Sentiment analysis on reviews is more favorable. Opinion mining can be implemented on difference levels of a sentence. It could be on word level, sentence level or document level. Most reviewers used to discuss over multiple subjects or aspects in a single review. If the sentence is analyzed as a whole it is unable to identify difference aspects which are mentioned by customers. Hence when the analysis is implemented on lower levels the opinions on various aspects can be identified separately. This process is known as aspect-based sentiment analysis. Many researchers are working on this area in present. This approach can be utilized for mining restaurant reviews in this research as well.

If search for literature on this area, several approaches could be found by various authors. In [1] aspect-based sentiment analysis is implemented using Naïve Bayes classifier. Main tasks are aspect detection and polarity detection towards an aspect category. Part of speech tagging is done using Stanford CoreNLP. Integrated lexical and rule-based approach for aspect-based sentiment analysis on government mobile app reviews domain is used in [2]. Aspect extraction, aspect sentiment scoring, Aspect sentiment aggregations are main tasks. Confusion metrics is used to measure the performance.

1.2 Aim

The aim of this research is to implement a reliable system for facilitating the searching of good restaurants in a particular city, providing a sentimental comparison over various restaurants features such as service, food, ambience and price etc.

1.3 Objectives

- Gathering valid customer reviews for several restaurants in a particular city from Yelp.com.

- Building and training multilabel Naïve Bayes model, SVM model, Convolution neural networks with pre-labeled restaurant datasets from International workshop on semantic evaluation in 2014.
- Preparing and preprocessing datasets.
- Extracting aspect terms and generating sentiment(positivity/Negativity) score values for aspect terms according to a rule-based algorithm applied on opinion words.
- Implementing word2vec trainings and finding cosine similarity between aspect terms and aspect categories in order to categorize aspect terms. If the assignment is failed using word2vec functions due to unable to match the threshold value, the trained machine learning model with highest accuracy should be used for category classification/prediction.
- Visualizing overall positivity or negativity towards each feature/aspect category using graphs.
- Comparison of restaurants in a particular city based on overall positive sentiment towards a feature.

1.4 Proposed Solution

Solution in this research is based on aspect-based sentiment analysis techniques. Both supervised and unsupervised techniques are used for the implementation. Machine learning techniques such as multilabel Naïve Bayes, SVM model and Convolution neural networks are trained, tested and compared for accuracy. Those models perform classification and prediction tasks of aspect categories. The methodologies used in the research are machine learning and rule-based approaches. Word embedding techniques such as word2vec model is trained with Google news dataset for detecting semantic similarities between aspect terms and aspect categories. The rule-based approach is used for assigning a sentiment score for aspect terms, based on positivity and negativity of the opinion words associated with those terms. The research focuses on providing a sentimental comparison over popular aspects/features of multiple restaurants such as food, price, ambience and service. Yelp.com is recognized as a user review collecting web site which helps public to find about local businesses through reviews. It does both recommending and advertising on businesses.

It also aids new businesses to make awareness among public about their products or services. This analysis work is based on yelp restaurant reviews.

In this report the chapter 2 represents an overview of literature review of related work based on natural language processing area. Chapter 3 describes the technology adapted to solve the problem and the chapter 4 includes methodology of the proposed system.
Chapter 5 contains the design and analysis of the proposed solution. Chapter 6 describes the implementation details of the system. Chapter 7 delivers a discussion with an evaluation of the results. Chapter 8 concludes the research.

Chapter 2

Literature Review on Natural Language Processing Systems

2.1 Introduction

Although natural language processing related tasks are known as complex, many researches are conducting various types of work related to this area. Each work consists of similar natural language approaches which are varying in methodologies. It is a branch of artificial intelligence and there are several sub categories which are capable of dealing with natural languages in different ways. Basic idea behind natural language processing is to identify features of human languages using computers and performing useful tasks. Sub areas such as syntaxes, sentiments, semantics, discourse, speech and dialogue are being researched in many ways. Different types of applications are implemented. Customer review mining is a recognized area as well. Reviews which are based on different domains are analyzed. Among them literature behind sentiment analysis-based researches are studied and compared with the improvements prior to perform this research. Although many research works are done based on sentiment analysis, very few are implemented on detecting specific aspects and making comparisons on them.

2.2 Related work based on Natural Language Processing

Aspect based sentiment analysis works accurately for identifying individual features of a sentience. This approach is found in many works with different methods. Aspect-based sentiment analysis is implemented using Naïve Bayes classifier in [1]. Main tasks are aspect detection and polarity detection towards an aspect category. Part of speech tagging is done using Stanford CoreNLP. Naïve Bayes classification is mentioned as a good approach for classification. Feature selection is done using Chi square.

Integrated lexical and rule-based approach for aspect-based sentiment analysis on government mobile app reviews domain is used in [2]. The lexicons are manually generated. Aspect extraction, aspect sentiment scoring of implicit and explicit sentiments, aspect sentiment aggregations are main tasks.

After extracting aspects, they are classified to find sentiments. confusion metrics is used to measure the performance. Student opinions mining is implemented for student recruitment and retention in [3] for higher educational institutes. It identifies the student satisfaction towards various aspects of educational institutes. The corpora used for this research is students' opinions about their courses, study programs and teaching staff. Aspect identification and sentimental polarity detection are main tasks. Machine leaning techniques with multi class classifications, term frequency inverse document frequency, Bag of word techniques and various n-grams are used together in [3] for analysis.

Star rating based solution for pre-labeled aspects is conducted in [4] using restaurant reviews. Aspects are extracted using Stanford Core NLP. Also, sentiment values are calculated using AFINN library. A word cloud is generated in order to visualize the frequent aspect terms. When a new term is added to the cloud it will be updated accordingly. Image of the word reveals the frequency of words by increasing the thickness of colors used to represent the word. The main research tasks in [4] are noun extraction, scoring algorithm implementation for sentiment values calculation and word cloud generation for visualization. Overall sentiment scores are translated in to star based rating systems.

In [5] analysis is done on specific restaurant aspects (price, service, food). The main steps of research are aspect term extraction, keyword extraction, aspect categorization and sentiment analysis. The author in [5] has used a topic modelling approach with double propagation method. Term frequency inverse cluster frequency and hybrid ELMO Wikipedia is proposed. And it is compared with other approaches such as wordvev2 and Glove. Name entity recognition is also performed. Sem - Eval 2016 dataset is used for classification. This dataset is manually labeled. Due to the imbalance of aspect categories author has done under sampling. Also, for sentiment analysis, expanded opinion lexical based method is used. In [5] the author has attempted to increase f1 value. In [6] machine learning based approach is used for ontology enhanced method. Two algorithms are implemented in order to perform tasks. A review-based algorithm and sentence aggregation method. Sentiment classification is done using training a multi class support vector machines.

Stanford Corenlp is used for various preprocessing tasks, and parsing. Restaurant ontology is described with three main classes as entity, property and sentiment. SVM model identify various features for classification. Feature generators and feature adapters. Author concludes that review-based algorithm delivers more accurate results than aggregated sentiment-based algorithm. Also, in [6] it is said that with ontology-based approach the quantity of training data does not make any sense.

For implementing sentiment analysis several classification techniques are used in [7]. Naïve Bayes, central based classifier, multilayer perception and support vector machines. In [8] the author states the importance of social media comments. It tries to extract useful information from you tube comments on videos. Machine learning based approach is used with naïve based classification. Author further states that using key words it is possible to identify sentiments of the user. The research focuses on comparative aspects. It adapts how users write comments comparing more than one product. Important tasks of research in [8] is class balancing and data preprocessing, classification and generating naïve Bayes probabilistic classifier. Weka is also used to get a deeper insights of classification model. It also states that weka does not support multi label classification. Mainly the author in [8] has performed Android and iOS comparison. Meka is used for multi label classification. Firstly, author has classified full comments and then in order to reduce computational power comments are filtered as nouns, adjectives and verbs. Also, author states that there is a naïve assumption for neighborhood keywords perform well.

In [9] the author states of sentiment analysis on Russian hotel reviews. Main tasks of the research are normalization, terms extraction and aspect score calculation. In terms extraction word vectors are generated using word2vec and cosine similarity measure is used for finding word and aspect category similarities. Also, word vectors are visualize using t-SNE algorithm. A weight value is calculated and for aspect terms and aggregated to find overall value. The results for implemented algorithm in [9] shows that SVM model performs well in classification tasks.

In [10] the author is using an aspect-based sentiment approach for SemEval-2014 datasets of laptops and restaurants. Also, it has used the foundation introduced by SemEval-2014 founders. This work also focuses on aspect category detection and polarity detection on aspect categories.

Techniques such as conditional random field is used for aspect extraction in the research. Also, [10] states the importance of using Z-score for aspect detection. Author has used it to identify aspect terms separately in aspect categories. Multinomial Naïve Bayes is used aspect polarity detection. Again SemEval - 2014 tasks are followed in [11]. Aspect category detection, opinion target expression, sentiment polarity classification are main tasks. Aspect category detection is performed using random forest classifiers. Sentiment polarity classification is implemented using bag of words and wordnet synset features. Several classifiers such as stochastic gradient descent, SVM, Adaboost are used for sentiment polarity detection.

Author in [12] has used yelp reviews for research. It is based on Latent Dirichlet allocation techniques. Also support vector machine with a fuzzy domain ontology is used. Data prepossessing step is performed as a usual step. Restaurant aspect extraction is implemented with Latent Dirichlet allocation. Sentiment detection is used with SentiStrength. Three types of evaluations are done in [12]. Single naïve Bayes model and then naïve Bayes model is used with support vector machine. Finally, the author has used a fuzzy ontology-based approach with support vector machine model.

In [13] a machine learning approach is used to extract information from the online reviews and find a feature-wise score and overall universal product score for each product using reviews. Naïve based classification is used. Reviews are analyzed and feature wise score is calculated. Datasets are preprocessed using noise removal, stop words and stemming. Extracted words are classified into positive and negatives using unigrams and Naïve Bayes classifiers. Machine learning based lexicon tool called sentiwordnet is used. The mechanism is based on bags of words generated from applying machine learning techniques. N-grams and multinomial Naïve Bayes model is used. Heavy weight bigrams are used.

Author has proposed a system to aspect and sentiment extraction and determination of sentiment orientation in [14]. Author propose a system to extract the aspect sentiment pair and compute the rating for each grouped aspect. The approach begins with selecting the subjective sentences in the review. Then it extracts aspects and opinions from the sentence and determine the orientation of the sentence.

A hotel review classification system using sentiment analysis is implemented in [16]. It is based on an existing corpus and a lexicon-based approach is used. The work consists of three different steps. First a corpus of lexicons is built with the components with semantic orientations. Next step is to perform sentiment analysis for classifications. Finally, classification results are evaluated against quantitative results. Also, two different setups are prepared to demonstrate the flexibility of the system.

In [17] the classification process is performed by several machine learning techniques. Naïve Bayes model, Support vector machines and decision trees are used. Evaluation of the model is performed by 10 fold cross validation. The author has noticed the increasing demand over smart phones. Hence the corpus used is unstructured dataset of mobile phones from Amazon.com. Dataset are made free of noisy data and preprocessed. Models are cross validated in order find the best classifier for the research.

Fine grained sentiment analysis process is proposed on [18]. It focuses on Semi structured reviews which are listed with pros and cons with short phrases. Each phrase is considered as segmented and these segments may contain aspects or the features. Unstructured reviews are composed of several sentences and there is no separation between pros and cons. Sentences are very long. Author has used these reviews considering that they are carrying useful information with more details. It states that handling such reviews are challenging comparing to handling short semi structured reviews. Extraction rules are created manually using parsers. They are applied for each subjective sentence. The process contains three steps as mentioned by the author. That is extracting product features and opinions, propagating product features and opinions and associating product features and opinions.

In [19] the approach is fully based on aspect-based sentiment analysis tasks presented on SemEval processes. It focuses on detecting aspect categories, finding sentiments towards aspect categories, detecting aspect terms and finding sentiment towards aspect terms. Dataset used in the research are restaurant s and laptops. An in-house sequence tagger and a supervised classifier to detect aspect categories, sentiment towards aspect terms and sentiment towards aspect categories is proposed in this work. Apart from the features such as n-grams, this approach benefits from using existing and newly created lexicon resources such as word -aspect association lexicons and sentiment lexicons.

A general process for sentiment polarity categorization is proposed in the research implemented in [20]. Datasets used in the process are online product reviews from Amazon.com. experiments for both sentence level categorizations and review level categorization id described with details. Several machine leaning techniques such as Naïve Bayes, support vector machines, Random forest are used in the research.

In [21] the authors have proposed novel unsupervised and domain independent model for detecting explicit and implicit reviews for sentiment analysis. The model is implemented in several steps and firstly a generalized method is proposed to learn multi-word aspects. And then a set of heuristic rules are defined, considering the influence of an opinion word for detecting the aspect. Secondly a metric based on mutual information and aspect frequency is proposed to score aspects with a new iterative algorithm. Thirdly a method for removing aspects considering the relations between aspects is implemented. Final model identifies implicit aspects using explicit aspects and opinion words.

The distance supervision techniques are used in [22]. Two main tasks are considered such as identifying relevant product aspects and determining or classifying sentiments. To different levels of granularity namely expressions vs sentence levels are considered. Dictionary based supervised approach and also several distant supervision techniques are used. Aspect detection at the expression levels is considered as a terminology extraction problem. At the sentence level it is considered as multi-label text categorization problem. The author has presented sentiment lexicon acquisition and sentiment polarity classification.

In [23], the implementation is based on an aspect-based sentiment analysis approach by modelling the interdependencies of sentences in the review using a hierarchical bidirectional LSTM. The author shows that the hierarchical model outperforms two nonhierarchical baselines to obtain results of competitive with the state of the art on five multilingual, multi- domain datasets.

Implemented system is based on aspect-based sentiment analysis in [24]. It consists of two components. Binary classifiers trained using single layer feedforward network aspect category classification and sequential labels classifiers for opinion target extraction.

The proposed method in [25] classifies subjective and objective reviews from blogs comments. The semantic core of subjective sentences is extracted from Sentiwordnet to calculate their polarity as positive, negative and neutral based on the contextual sentence structure. It uses machine learning techniques as well.

Above mentioned researches are implemented on aspect-based sentiment analysis. With the support of literature, it is proved that review mining is important in present. Studying related works helps to gain knowledge in a particular area and aids to identify the correct path for continuing the research further. This research is also based on aspect-based sentiment analysis techniques for restaurant reviews. Most of the previous researches are based on detecting particular aspects. Comparison based researches are less. Hence this work focuses on comparing features of multiple restaurants. A hybrid methodology will be used.

Next chapter presents the suitable technologies for the development of the research and adaptation of those technologies for solving the research problem.

Chapter 3

Automation of Restaurant Feature Comparison Based on Customer Experience

3.1 Introduction

Sentiment analysis is a recognized research area. It is a sub category of natural language processing. New applications related to natural languages are emerging continuously in this area. With the extensive usage of internet applications such as social media, forums, blogs, e commerce web sites such as Amazon.com, TripAdvisor.com, people are sharing their options based on real experiences. These comments have sentiments of customers. It affects online business operations. Operation teams in business organizations have to listen to the customers voice if they need to make improvements. Customer reviews are huge in number. Reading them is a difficult task to perform. Hence review mining has become popular. By extracting the sentiments of the customer reviews within a short period improves speedy decision making over customer satisfaction. There are several aspects of a business to take care of. Customers also have different views towards different aspects of a business. They write comments mentioning those aspects separately. Hence review mining should be done for individual aspects. It makes business people to realize which areas should have more attention. Solutions based on sentiment analysis can resolve above problem.

Sentiment analysis tries to detect sentiments (positivity or negativity) of sentences. Aspect-based sentiment analysis can be used to perform word level analysis of customer reviews. Reviewers mention various aspects in a single comment/sentence. Also, these aspects contain different opinions. One aspect can have a positive opinion while other aspect has a negative opinion. Hence classifying such reviews is a complex task. While studying about the literature regarding this topic, different implementation approaches are found. Although each research contains similar processes for basic natural language processing tasks such as preprocessing for dataset preparations, other analysis tasks are performed using different approaches.

Also, the researchers are based on different domains that need different solutions. Appearance of the dataset varies with the domain and implementation methodologies are varying accordingly. Most of the technologies used in the solution area can be either machine leaning based supervised techniques or unsupervised techniques. Several common steps are conducted in aspect-based sentiment analysis researches [9,10,11]. Aspect terms extraction, aspect category detection, polarity detection of aspect categories is among them. Researches have used different methodologies for implementing these steps. But a comparison of multiple restaurant features is rare in the literature. This research focuses of implementing a comparison system. Following technologies are considered useful for the development of this research.

3.2 Natural Language Processing Techniques

Basically, natural language processing is about making computers to understand human languages in order to perform various tasks. [15] states that it is a subfield of computer science, linguistics and artificial intelligence connected with the interaction of human languages and computers. Important tasks of the field are speech, syntaxes, semantics, discourse and dialogue. In early stages hand coding rules were used for implementations and then machine learning techniques were introduced. These approaches automatically learn rules of a large corpus. It made improvements in this field. Several learning model types are available such as neural networks decision trees, support vector machines, Regression analysis, genetic algorithms and Bayesian models. There are different types of learning algorithms as well. Supervised learning, unsupervised learning, semi supervised learning, self-learning ,reinforcement learning and feature learning etc. Algorithm will be selected according to the inputs, outputs and the type of the solution needed. This research also uses machine learning techniques such as Naïve Bayes model, support vector machines and convolution neural network with deep learning in order to train large corpus of restaurant reviews. Most accurate model is used for further predictions.

3.3 Sentiment Analysis

It is all about identifying the polarity/sentiment towards texts in a document, sentence or an aspect. In [15] it is interpreted as a subfield of natural language processing and computational linguistics which does the identification and extraction of the subjective information from texts. Most of the events it is used in marketing evaluations which identifies the voice of customer. This can have many forms such as aspect-based sentiment analysis, grading sentiment analysis, multilingual sentiment analysis. In [15] it is also said that customer review mining is implemented mostly with the aid of sentiment analysis. Several machine learning techniques are used for review mining such as support vector machines, naïve Bayes or deep learning. Following figure shows the summarization of sentiment analysis techniques.

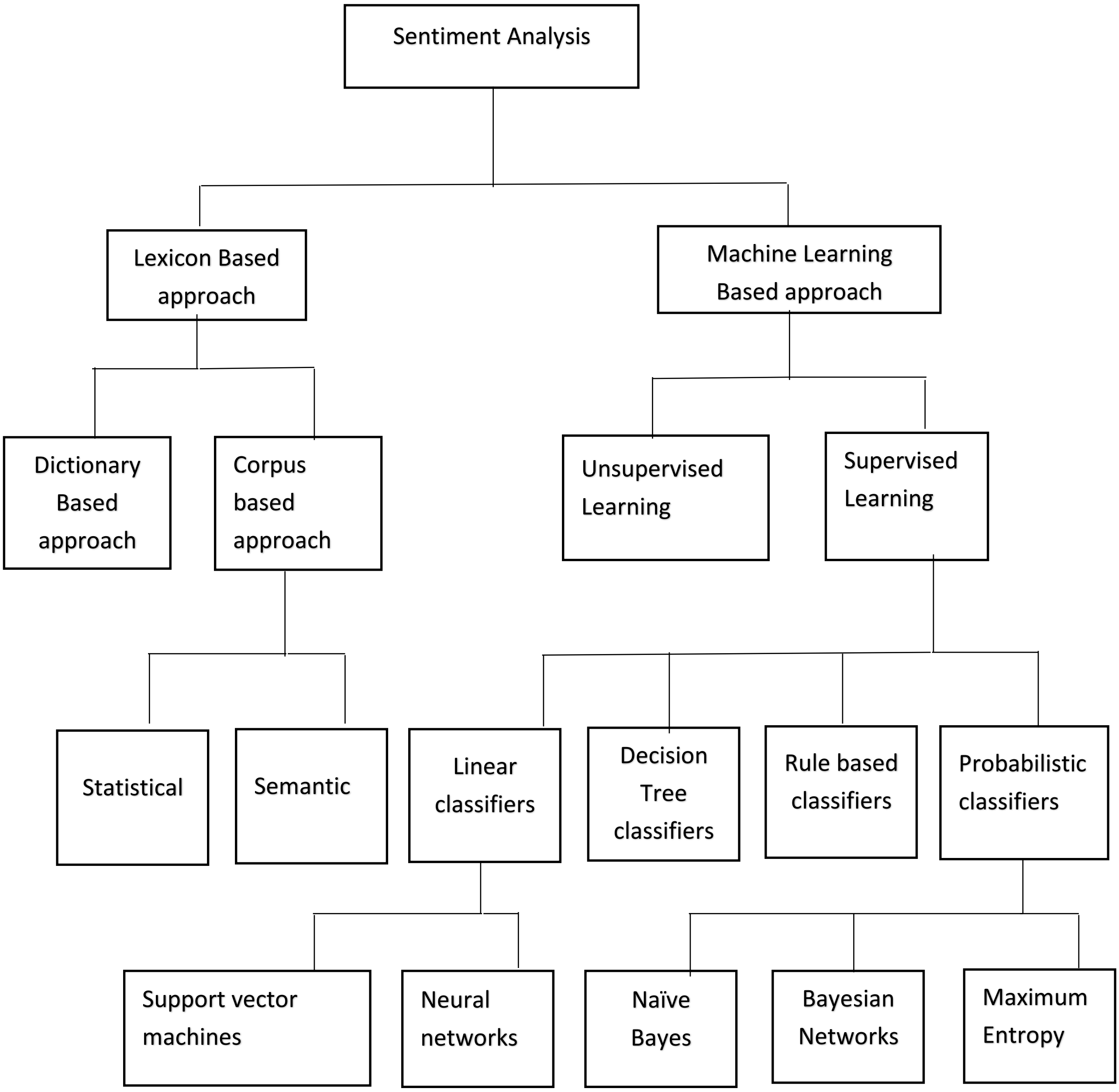


Figure 3.1: Sentiment Analysis Techniques Summary

3.4 Feature/Aspect based sentiment analysis

This is about discovering sentiments towards aspects related to a specific entity. Same text may contain information about different entities. Opinions towards entities may also different. Ex: “The food is very tasty but the waiter was not very helpful”. In this example there are two aspects are being considered about. Food and service. Sentiments towards food is positive but service gives a negative opinion. This kind of varieties should be handled individually. When performing this kind of analysis first need to identify the relevant entities, extracting their features and classifying the polarities. Feature identification can be done by using methods such as topic modeling or deep learning.

This chapter presented suitable technologies that can be utilized in order to solve this research problem. Next chapter will deliver the methods of adapting these techniques correctly for solving the problem.

Chapter 4

Automated system for Feature Comparison of Yelp Restaurants

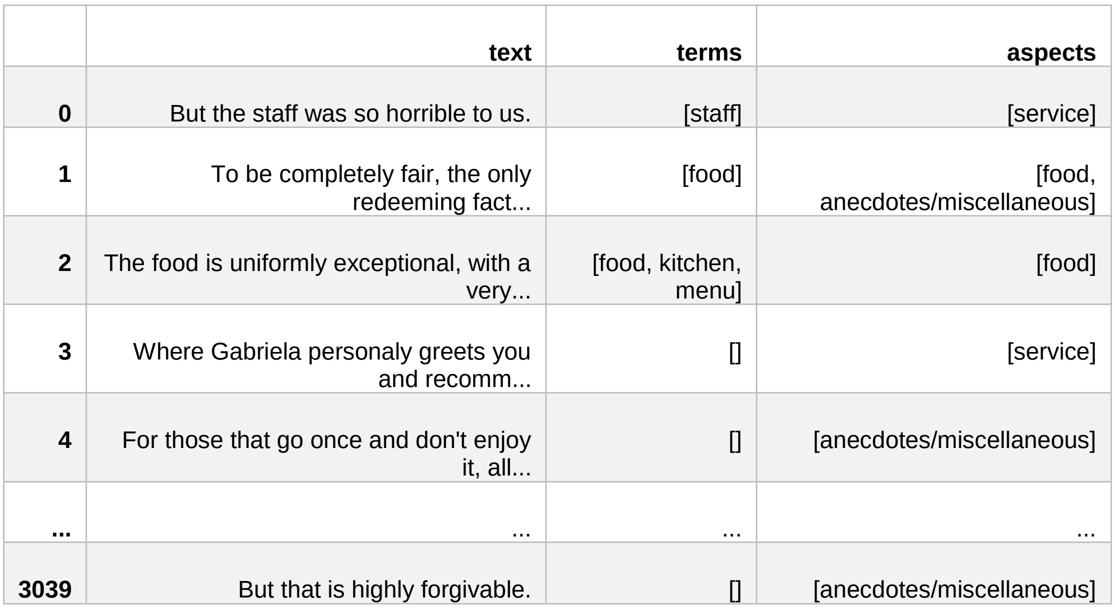
4.1 Introduction

This research is based on implementing a restaurant review mining system using feature/aspect-based sentiment analysis. There are popular restaurant features which are being criticized or praised by visitors of the restaurants such as food, prices, services and environment. Most of the previous researches performed on this domain are based on these features. This analysis work focuses on comparing these features of multiple restaurants. Placing restaurant reservations online is very popular in present. Before making reservations, people try to get others opinions or recommendations about the restaurant. Also, they give attention for particular aspects of the restaurants. But reading customer comments on multiple restaurants and searching about each feature is a time-consuming task. Hence automating the comparison of restaurant features can be a requirement for users as well as for the staff. Business owners can give attention to customer voice immediately and take necessary actions to make further improvements. This can be done with mining customer reviews and extracting useful information from these reviews. User reviews from Yelp.com is used in this research. Yelp is a website which collects reviews from users about local businesses. It gives space for users to write comments. Anybody can read this review in order to get an idea about the business. Yelp datasets of restaurants are gathered in order to classify. The research contains multiple steps such as dataset preparation and preprocessing, machine learning model building, trainings and testing, aspect term extraction, sentiment polarity detection/sentiment score calculation for those aspect terms and categorization of terms into separate categories. Finally, most frequent positive or negative terms of categories are detected and overall positive values for aspect categories are visualized and compared.

4.2 Dataset preparing and preprocessing

Data pre-processing is about transforming incomplete, inconsistent data into readable and understandable formats. Row data is converted in to more usable formats. Several restaurant datasets were used in this research such as SemEval-2014 datasets and multiple restaurant datasets from Yelp.com. Former datasets are from International Workshop on Semantic Evaluation (SemEval) in year 2014. The datasets contain predefined aspects, terms and text/review. SemEval – 2014 consists of four aspect categories such as food, price, ambience, service and miscellaneous. Table 4.1 depicts the SemEval-2014 Dataset. Datasets are split into two types as test dataset and training dataset. Several preprocessing tasks were performed over the datasets in order to perform the analysis task efficiently.

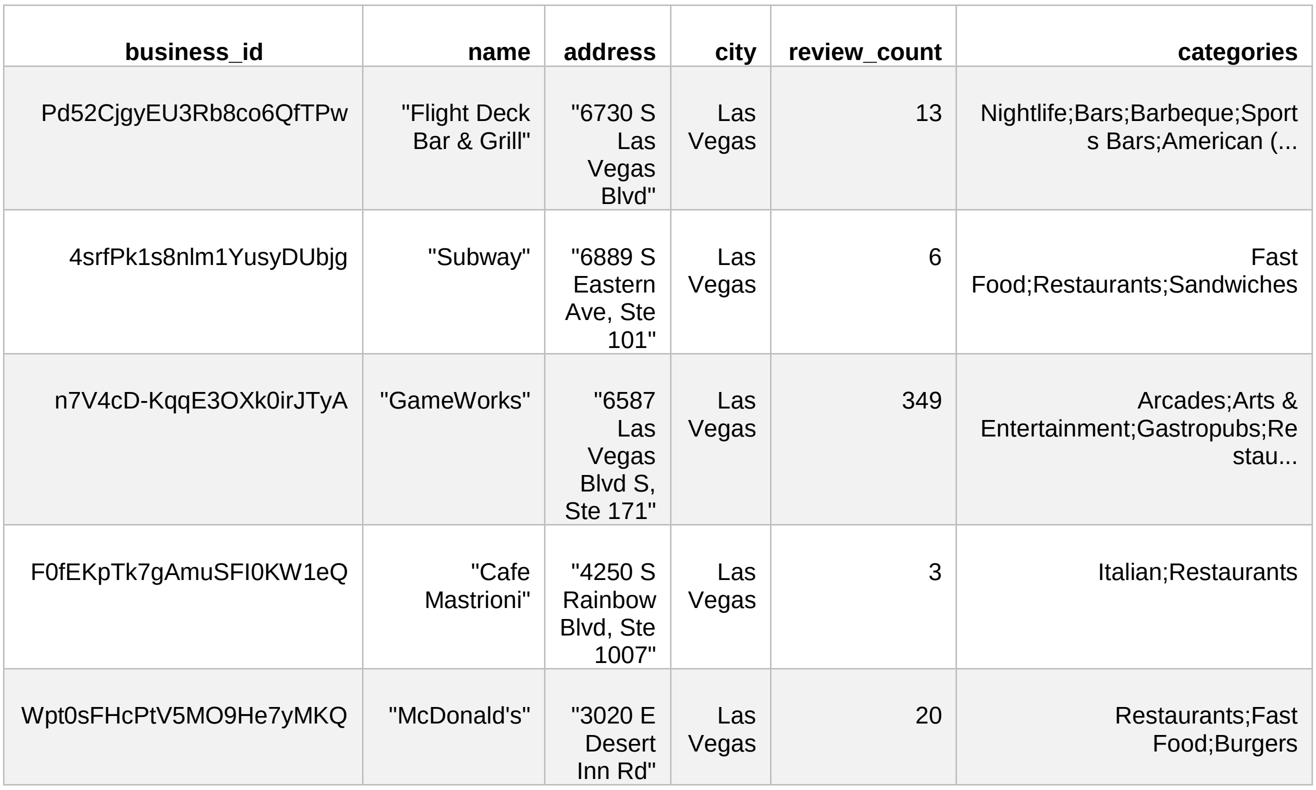
From the following dataset 'Text' and 'aspects' attributes are selected for training the machine learning models. The 'text' attribute contains customer review and preprocessed with lowercase, stop words, tokenization methods along with vectorization methods such as count vectors, term frequency (Tf-idf) and Bag-of-words in deep learning.



	text	**terms**	**aspects**
0	But the staff was so horrible to us.	[staff]	[service]
1	To be completely fair, the only redeeming fact...	[food]	[food, anecdotes/miscellaneous]
2	The food is uniformly exceptional, with a very...	[food, kitchen, menu]	[food]
3	Where Gabriela personaly greets you and recomm...	[]	[service]
4	For those that go once and don't enjoy it, all...	[]	[anecdotes/miscellaneous]
...	...	...	...
3039	But that is highly forgivable.	[]	[anecdotes/miscellaneous]

Table 4.1: SemEval-2014 Dataset

Two separate datasets from Yelp were important for this work. Those were business details dataset and customer review details dataset. Table 4.2 contains the business details dataset. It contains attributes such as Business id, Business name, Address, City, Review count and Business category.



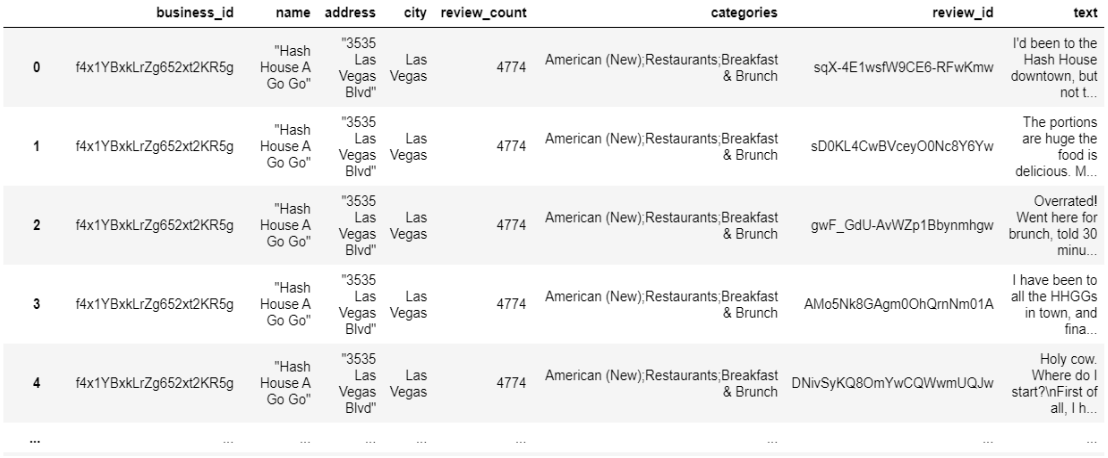
business_id	name	address	city	review_count	categories
Pd52CjgyEU3Rb8co6QfTPw	"Flight Deck Bar & Grill"	"6730 S Las Vegas Blvd"	Las Vegas	13	Nightlife;Bars;Barbeque;Sports Bars;American (...
4srfPk1s8nlm1YusyDUbjg	"Subway"	"6889 S Eastern Ave, Ste 101"	Las Vegas	6	Fast Food;Restaurants;Sandwiches
n7V4cD-KqqE3OXk0irJTyA	"GameWorks"	"6587 Las Vegas Blvd S, Ste 171"	Las Vegas	349	Arcades;Arts & Entertainment;Gastropubs;Restau...
F0fEKpTk7gAmuSFI0KW1eQ	"Cafe Mastrioni"	"4250 S Rainbow Blvd, Ste 1007"	Las Vegas	3	Italian;Restaurants
Wpt0sFHcPtV5MO9He7yMKQ	"McDonald's"	"3020 E Desert Inn Rd"	Las Vegas	20	Restaurants;Fast Food;Burgers

Table 4.2: Yelp Business Details Dataset

Table 4.3 contains the customer reviews dataset which contains review id, business id and customer review /text. These two were filtered and merged over restaurant business category using the business id. Also, restaurants in same city (Las Vagas) were selected. Figure 4.1 depicts the merged dataset. Final dataset consists of Business Id, Business name, Business category, Business address, City, Review, Review Id and Review count.

	review_id	business_id	text
0	vkVSCC7xljjrAI4UGfnKEQ	AEx2SYEUJmTxVVB18LlCwA	Super simple place but amazing nonetheless. It...
1	n6QzIUObkYshz4dz2QRJTw	VR6GpWIda3SfvPC-lg9H3w	Small unassuming place that changes their menu...
2	MV3CcKScW05u5LVfF6ok0g	CKC0-MOWMqoeWf6s-szl8g	Lester's is located in a beautiful neighborhoo...
3	IXvOzsEMYtiJI0CARmj77Q	ACFtxLv8pGrrxMm6EgjreA	Love coming here. Yes the place always needs t...
4	L_9BTb55X0GDtThi6GlZ6w	s2I_Ni76bjJNK9yG60iD-Q	Had their chocolate almond croissant and it wa...
...	...	...	...
999995	MJfGu0-OYvl3_VMOZwSTEw	Ec8fKdO0oxZNOCeG2ThwpQ	This is the best vape shop I've been to. Every.

Table 4.3: Yelp Customer Review Details Dataset



	business_id	name	address	city	review_count	categories	review_id	text
0	f4x1YBxkLrZg652xt2KR5g	"Hash House A Go Go"	"3535 Las Vegas Blvd"	Las Vegas	4774	American (New);Restaurants;Breakfast & Brunch	sqX-4E1wsfW9CE6-RFwKmw	I'd been to the Hash House downtown, but not t...
1	f4x1YBxkLrZg652xt2KR5g	"Hash House A Go Go"	"3535 Las Vegas Blvd"	Las Vegas	4774	American (New);Restaurants;Breakfast & Brunch	sD0KL4CwBVceyO0Nc8Y6Yw	The portions are huge the food is delicious. M...
2	f4x1YBxkLrZg652xt2KR5g	"Hash House A Go Go"	"3535 Las Vegas Blvd"	Las Vegas	4774	American (New);Restaurants;Breakfast & Brunch	gwF_GdU-AvWZp1Bbynmhgw	Overrated! Went here for brunch, told 30 minu...
3	f4x1YBxkLrZg652xt2KR5g	"Hash House A Go Go"	"3535 Las Vegas Blvd"	Las Vegas	4774	American (New);Restaurants;Breakfast & Brunch	AMo5Nk8GAgm0OhQrnNm01A	I have been to all the HHGGs in town, and fina...
4	f4x1YBxkLrZg652xt2KR5g	"Hash House A Go Go"	"3535 Las Vegas Blvd"	Las Vegas	4774	American (New);Restaurants;Breakfast & Brunch	DNivSyKQ8OmYwCQWwmUQJw	Holy cow. Where do I start?\nFirst of all, I h...
...	...	...	...	...	...	...	...	...

Figure 4.1: Yelp Review Details Dataset

From the above dataset the ‘text’ attribute contains the customer review. Each text was split into single sentences and lower case, tokenization, stop words and removal of special characters were performed in order to prepare and preprocess the data.

4.2.1 Lowercase

Dataset contains the same word with lowercase or uppercase. This may deliver different output. Hence in order to preserve consistency this simple step is important in nlp tasks.

4.2.2 Stop words and Noise Removal

Stop words are known as the common words used in a language such as 'a, 'the', 'is', 'are' etc. Analysis tasks can be done with eliminating these words and focusing on important words. Sentences are also containing special characters, digits and other unnecessary text pieces. These things make noises for analysis tasks and should be removed.

4.2.3 Tokenization

Longer strings of texts should be split in to pieces. If the text contains multiple sentences it should be split into single sentence.

The sentences should be tokenzed in to words/tokens or chunks. The analysis processes are performed upon these chunks. This process is also called as text segmentation.

4.3 Machine Learning Approach for Review Classification

Text/review classification is an automated process of classification of texts into predefined categories. There are several machine learning techniques which can be used for text mining. With the recent advancements in machine learning techniques such models are capable of learning a large corpus automatically. This research work is implemented using machine learning techniques such as Naïve Bayes model, Support vector machine and deep learning methods with Convolution neural networks which are very competitive over each other. They made computers to understand complex human languages with lot of ambiguity which does not follow any logical rules.

These models are to be trained and tested by annotated restaurant reviews of SemEval -2014. Reviews are manually labeled with aspect categories. Dataset is divided into training and testing datasets. These models will be used to classify the aspect category for the terms included in yelp review sentences.

This approach is used when aspect terms are categorizing into aspect categories. Following sub sections provide details on classification processes using machine learning models.

4.3.1 Naïve Bayes Classification

This is a simple classifier based on probabilities of events. Training datasets required for each categories (classes). Model can be trained to classify customer reviews for relevant categories. SemEval -2014 dataset contains reviews for several categories such as food, price, ambience and service. Some reviews cannot be categorized into a relevant restaurant aspect. Hence those are categorized as miscellaneous.

4.3.2 Naïve Bayes Model

Naïve Bayes model is often used by researches in sentiment analysis. It is a probabilistic model which helps for classification tasks which is based on Bayes theorem. Depends on probabilities of events. It remains popular for text classification tasks and delivers accurate, reliable outputs.

Model helps to find conditional probabilities of occurrence of two events based on the probabilities of occurrence of each individual event. Probabilities are computed over training sets.

It computes the probability of categories for each term and outputs the category with highest probability. It uses Bayes theorem which uses the probability of feature based on prior knowledge of conditions related to that feature.

Bayes Theorem -

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Multinomial Naïve Bayes model is used in this research. It is a special instance of a naïve bays classifier which uses a multinomial distribution for each of the feature. ML Naïve Bayes model is trained for classifying restaurant aspect categories.

Word frequencies are used as features for the model learning. Review sentences should be transformed in to vectors before feeding into the model. Individual words are considered as independent of sentences and word frequencies are used for calculating probabilities.

Advanced techniques such as stop words, n-grams, Term - Frequency Inverse Document Frequency, Count vectorizer are used in order to develop the model more efficiently. The process is called feature engineering. The data is converted in to feature vectors and new features are generated using existing data.

- Count Vectors

It gives a matrix for each category in the dataset. This is known as Term Document Matrix. Contains the frequency of a word in the sentence Each row represents a document of the corpus. Each column is representing a term in a corpus. Also, each cell contains the frequency count of a particular term in a particular document.

- N – grams

Uses sequences of words instead of counting single words. Hence a token can be considered as a sequence of n items. Simple case is unigrams (1-gram) which uses a single word. Choosing the number of n depends on the language and the type of the application. This model uses bigrams.

- Term Frequency – Inverse Document Frequency

It represents how important a word is to a document. Often used in text mining applications. It contains a score. The score has two terms. One is a normalized frequency.

Second one is the inverse document frequency which is computed as the logarithm of the number of the documents in the corpus divided by the number of the documents where the specific terms were appeared.

- Label Powerset Transformation

This classification contains datapoints which can have multiple classes. Hence the problem is transformed into a multiclass problem with one multiclass classifier which is trained on all unique label combinations found in the training data. The method maps each combination to an id number and perform multi class classification using the classifier as multi class classifier and combination ids as classes.

4.3.3 Support Vector Machines

This is a supervised learning model which is also used in both classification and regression analysis. It is known to be using less computational power. It determines the best decision boundary between vectors that belong to a given category and vectors that do not belongs to it. Texts should be vectorized.

It reflects samples as points in space mapped so that the samples of separate categories are divided by a clear gap that is wide as possible. New samples also mapped into the same space and predicted to belong to a category based on the side of the gap on which they fall. SVM can efficiently represents a non-linear classification as well. This can be used with any kind of vectors which encode any kind of data. When classifying text representations, they can identify which side of the boundary they fall into. The model consists of advanced techniques such as stop words, n-grams, Term - Frequency Inverse Document Frequency and Count vectors. This process is called feature engineering. The data is converted in to feature vectors and new features are generated for further analysis.

4.3.4 Deep Learning Methods

This a sub field of machine learning and uses neural network architectures for classification tasks. They follow the human brains in processing data. These models can learn from unstructured data as well. A deep learning model is also trained with a suitable architecture and parameters for aspect category prediction in this research. Convolution neural network is a type of dep learning technique which is commonly used with computer vision related applications.

But recently it is being used with text classifications as well. Network architectures contain neurons organized in to multiple layers. Neurons in both layers are connected with each other.

Input layer is designed to get inputs and results are for the output layer. When the sample observations are given the model learns adjusting weights in order to increase the accuracy. This should be implemented by minimizing the errors. This is designed for minimal processing with multilayered perceptron. Model can detect patterns of multiple sizes such as n-grams. They could be expressions.

This CNN model is designed with four layers and trained with training data of review texts and aspect categories. Vector representation of words is Bag-of-words.

- Bag-of-words

It involves in extracting features from texts for using machine learning models. It is a representation of a sentence which describes the occurrence of words within a document consists of two aspects. One is the vocabulary of known words. The other is the measure of the presence of known words.

The model only checks whether the known words are occurred in the documents. Word count is considered as the feature. In this work, for each observation words are tokenized and finds the frequency of each token.

4.4 Rule based Approach for Aspect Term Extraction and Polarity Detection

Customer reviews contain several information about various aspects of a restaurant. A single review may contain information on multiple aspects such as food, service, ambience or price. It may also discuss about single aspect. Customers use various terms/words to mention above aspects in their reviews. For an example if a customer discusses about service aspect he/she may write “The waiter was not helpful”. Hence the exact nouns should be extracted in order to identify the subjectivity of the review. These nouns can be used as aspect terms which indicate the aspect categories of a restaurant. Some sentences may be completely anecdotal which delivers none of the useful information. A review can be positive or negative. Customers expose positive or negative ideas or opinions towards restaurant aspects through their reviews. If they are satisfied with a certain aspect they may write a positive comment or if unsatisfied the comment will be a negative one. For an example if a customer is unsatisfied and needs to deliver a negative comment on food the comment will be “meal was not good”. The positivity or the negativity of a review should be determined.

Dependency parser of spacy can be used to extract the aspect terms or the nouns of a review/sentence. It may help to reveal the subjectivity of the sentence. It is used to recognize the dependency links of relevant aspect terms and their correspondent opinion.

These opinion words help to identify the negativity or the positivity towards relevant aspect term. It is performed using opinion lexicons. Opinion words of each Yelp review sentence is checked by opinion lexicons in order to identify the polarity. Polarity can be positive or negative. A score will be given based on the polarity of an opinion word. If an opinion word is negative a low score is assigned to aspect term and high score is assigned if the opinion word is positive. Following topic gives a brief description on dependency parser which is used to perform aspect extraction and dependency link identification between aspect terms and opinion words.

4.4.1 Dependency Parser

This is a process of identifying a sentence and the grammatical structure. Also, it defines the relationship between head words and, words which modify those heads. The commonly used syntactical structure is a parse tree which is generated using parsing algorithms. It converts a sentence into a dependency tree.

Term extraction and identification of opinion words included in a sentence can be recognized using the dependency relationships. Assignment of sentiment scores for opinion words is a rule based process. Negation words are assigned in a different way based on those rules.

4.5 Aspect Categorization

Aspect term extraction and polarity detection is described in the previous topic. Extracted terms should be assigned into broader aspect categories such as price, food, service, and ambience. If the sentence is unable to categorize in to any of the above categories it is assigned in to miscellaneous category. This task can be performed using two methods. One method is calculating a value for semantic similarity between the extracted terms and the aspect category. Second method is by using the trained machine learning models. Calculation of similarity value is performed using word embedding methods. A word2vec model is trained using Google news dataset. Word2vec's n_similarity function is used to calculate similarity value between extracted term and aspect category. A threshold value is maintained in order to detect a good similarity score and to perform the assignment efficiently.

If the calculated value does not reach to the threshold value the assignment is performed using the trained machine learning classification model. Model with the highest accuracy can be used for classification. The entire review can be classified into one of the broader aspect categories using the rained model. Following topic gives a brief description on techniques used for aspect categorization.

4.5.1 Word Embedding Techniques

This is a method of representing words as vectors. Each word has a vector. Similar words may have similar vectors. Vectors are numerical representations. Hence a computer can handle them easily. There are various types of these embeddings. frequency based embedding and prediction based embeddings. Count vectors, TF-IDF, Co-Occurrence matrices are frequency based and bag of words, skip – Gram models are prediction based. Pre-trained word vectors are available as well as user defined vectors. Word2vec is a prominent word embedding technique. This research uses this algorithm for finding semantic similarities. This is useful for finding similarity between terms and aspect categories. Although there are several Pre- defined word vectors, dataset used in this work is has 3 million of vocabulary words trained on 10 billion of words from google news dataset.

- **Word2vec**

It is a word embedding technique for generating word vector representations for an input of text corpus. It gets the input as a text corpus and outputs are set of vectors. It has a distributed representation of word vectors.

It generates computer readable content. It is used in this research and is trained with Google’s news dataset. Word2vec functions are used for similarity detections between two words.

This chapter presents the way of using relevant techniques for the development of the research. The next chapter is representing the analysis tasks and designing of the system.

Chapter 5

Analysis and Design of Restaurant Reviews Mining System

5.1 Introduction

This research is about implementing a review mining system for restaurants. It is the aim. The system is designed with natural language processing techniques such as sentiment analysis and semantic analysis. Yelp.com is the provider of review datasets for generating the text corpus. It contains user reviews of several types of local businesses. Restaurant category reviews are extracted for this work. After studying and analyzing about various machine learning methodologies, three types of learning models are trained and tested for achieving results. Machine learning models are trained with SemEval-2014 dataset with pre-defined categories for aspect terms. Multinomial Naïve Bayes classifier, support vector machines and deep learning methods are used. The model with higher accuracy is used for further analysis. Tasks related to sentiment analysis and semantic analysis of review terms should be analyzed and designed prior to the implementation.

System should deliver a comparison of multiple restaurant over various aspects such as food, service, price and environment or ambience. The comparison system is designed with multiple steps. Mainly the sentiment orientation towards aspect categories should be detected and the comparison of those aspects over multiple restaurants should be visualized in order to perform easy decision making over the best restaurant. Firstly, the machine learning models are to be trained and tested with SemEval-2014 annotated review datasets for text classification. Then the aspect terms should be extracted and polarities should be detected for each aspect term. This is done using opinion lexicons. Dictionaries are generated with terms and sentiment scores. Extracted terms are to be compared for semantic similarities with broader aspects and categorized separately. Finally, the sentiment orientation towards aspect categories should be generated as positively and negatively. Following figure shows the top-level designing steps of the comparison system.

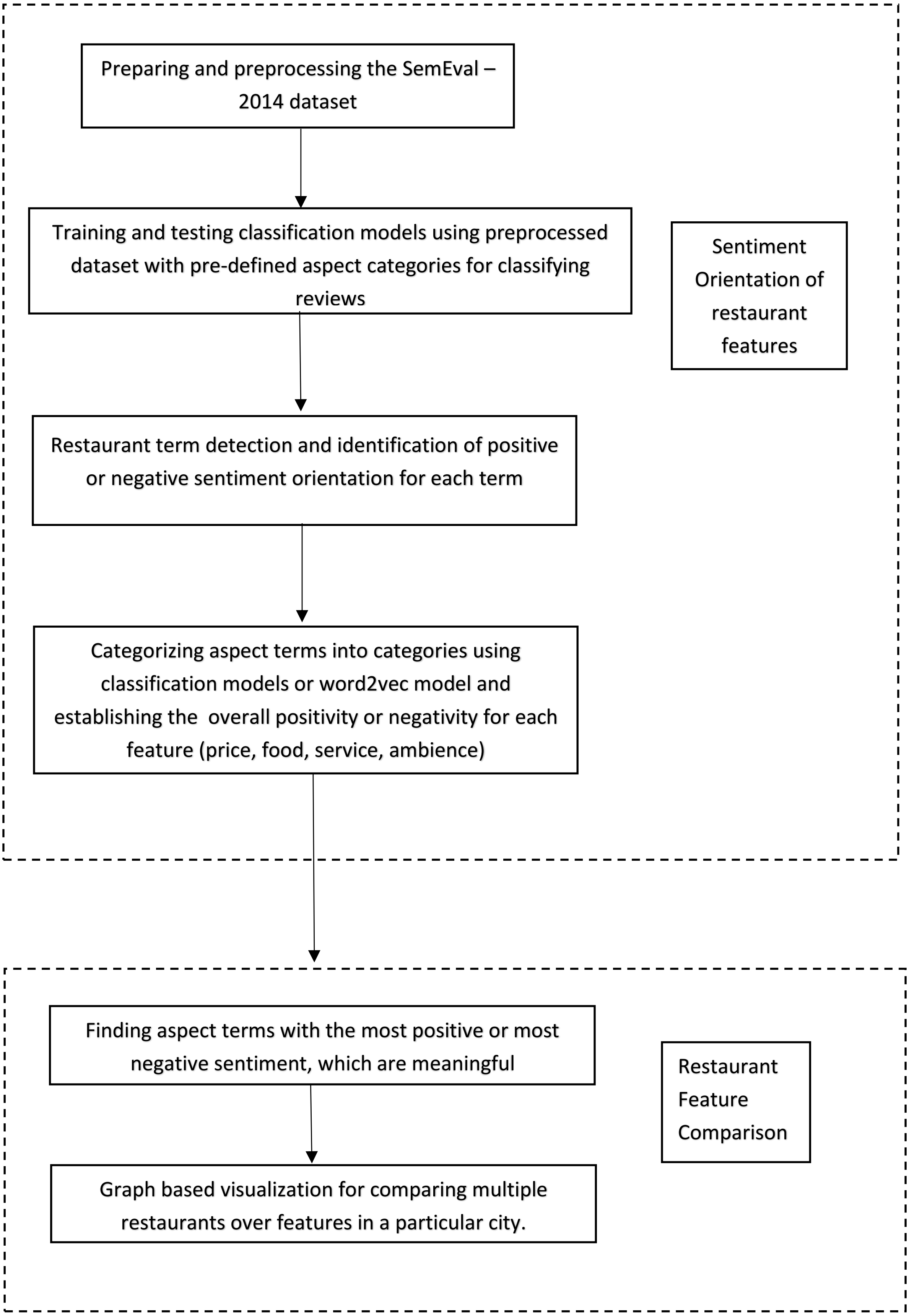

Preparing and preprocessing the SemEval – 2014 dataset
Training and testing classification models using preprocessed dataset with pre-defined aspect categories for classifying reviews
Sentiment Orientation of restaurant features
Restaurant term detection and identification of positive or negative sentiment orientation for each term
Categorizing aspect terms into categories using classification models or word2vec model and establishing the overall positivity or negativity for each feature (price, food, service, ambience)
Finding aspect terms with the most positive or most negative sentiment, which are meaningful
Restaurant Feature Comparison
Graph based visualization for comparing multiple restaurants over features in a particular city.

Figure 5.1: Top Level Design of the system

5.2 Text/Review Classification

Machine learning techniques such as Multinomial Naïve Bayes classifiers, Support Vector Machines and deep learning Convolution neural network classification methods are used. The model with highest accuracy value is used for further analysis. Models are trained with pre-defined aspects categories with reviews.

5.3 Extracting terms of Yelp reviews and generating dictionary with polarity for each term.

Terms extraction can be done with techniques such as syntactic dependency parsing. Each review is passed through a parsing tree and it identifies the grammatical patterns between words. These review sentences are tokenized and aspect terms are extracted. Opinion words are detected according to dependency links between terms and opinion words. In the next phase opinion lexicons are used to identify the positivity or the negativity of opinion words. A rule-based approach is used to detect the polarity of the opinion words. It is performed by calculating a score for each opinion word. A low score is given to negative opinions and a high score is given to positive opinions. Finally, a dictionary with score value and aspect term are generated for each review.

5.4 Assigning aspect terms in to aspect categories and detecting the overall sentiment (positive/negativity) towards aspect categories.

Aspect categories should be pre-defined. Food, price, ambience and service is used as aspect categories. For the assigning process word2vec model is trained using Google News Dataset and similarity values are calculated between extracted terms and broader aspect categories. Word2vec's n_similarity function is used for calculating the similarity. If the assignment is failed due to not reaching to the threshold value, the trained machine learning model can be used for category detection process. In the end each aspect category must find the overall sentiment towards positivity and negativity. It is performed by using the calculated polarity score value for each aspect term in the previous phase.

5.5 Finding most positive or most negative aspect terms of a category

This is useful for business owners. They can identify the frequent terms which are being discussed by customers often and take necessary actions for customer satisfaction. Dictionaries generated from dependency parsers contains terms with polarity values. Terms with highest values are considered as frequent terms.

5.6 Visualizing Aspect categories of multiple restaurants for decision making.

Final task is to calculate a total value for the positivity of each aspect category. These categories are to be visualized by a graph in order to make comparisons among restaurants.

This chapter delivers descriptions about designing and analyzing steps with suitable technology applications. Also, a high-level design of the system is also provided. The next chapter is providing details about implementation steps and used techniques for the development of the system.

Chapter 6

Implementation of Restaurant Feature Comparison System

6.1 Introduction

The implementation of the system consists of multiple steps. In the following sections, the necessary steps are explained in detail and the tools and technologies utilized for implementation are mentioned. Firstly, the datasets should be collected from Yelp.com.

6.2 System Development

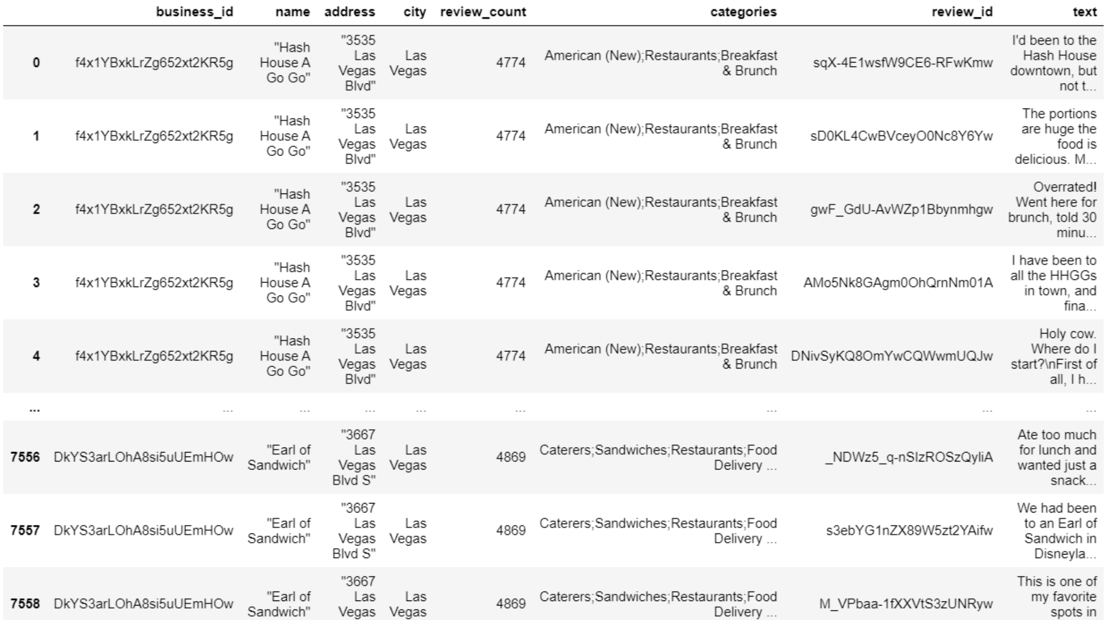
The system is implemented using Python3 and several data science tools. Other system requirements are as follows.

- CPU intel® core™ i7
- RAM 8.00 GB
- 64-Bit Operating system x64 based processor - Windows 10
- Python libraries - Genism, Spacy, Keras, Scikit-learn
- SemEval - 2014 dataset properties -
    - Aspect category
    - Customer review text
    - Aspect terms
- Yelp Dataset properties -
    - Business id
    - Business name
    - Business category
    - Business address
    - City
    - Review id
    - Review count
    - Review

6.3 Dataset preparing and preprocessing

Row data is converted in to more usable formats. Restaurant datasets were used in this research such as SemEval-2014 datasets and multiple restaurant datasets from Yelp.com. The datasets contain predefined aspect categories towards reviews. SemEval – 2014 consists of four aspect categories such as food, price, ambience, service and miscellaneous. Datasets are split into two types as test dataset and training dataset. Several preprocessing tasks were performed over the datasets in order to perform the analysis task efficiently. The dataset contains attributes such as Text/Customer review, terms and aspects. The 'text' attribute contains customer review and preprocessed with lowercase, stop words and tokenization methods together with vectorization methods such as count vectors, term frequency (Tf-idf) and Bag-of-words in deep learning.

Two datasets from Yelp.com are collected for this work. Those are business details dataset and customer review details dataset. These two are filtered and merged over restaurant business category using the business id. Also, restaurants in same city are selected. Further details about datasets were presented in chapter 4. Final dataset consists of Business Id, Business name, Business category, Business address, City, Review Id, Review count and Review



	business_id	name	address	city	review_count	categories	review_id	text
0	f4x1YBxkLrZg652xt2KR5g	"Hash House A Go Go"	"3535 Las Vegas Blvd"	Las Vegas	4774	American (New);Restaurants;Breakfast & Brunch	sqX-4E1wsfW9CE6-RFwKmw	I'd been to the Hash House downtown, but not t...
1	f4x1YBxkLrZg652xt2KR5g	"Hash House A Go Go"	"3535 Las Vegas Blvd"	Las Vegas	4774	American (New);Restaurants;Breakfast & Brunch	sD0KL4CwBVceyO0Nc8Y6Yw	The portions are huge the food is delicious. M...
2	f4x1YBxkLrZg652xt2KR5g	"Hash House A Go Go"	"3535 Las Vegas Blvd"	Las Vegas	4774	American (New);Restaurants;Breakfast & Brunch	gwF_GdU-AvWZp1Bbynmhgw	Overrated! Went here for brunch, told 30 minu...
3	f4x1YBxkLrZg652xt2KR5g	"Hash House A Go Go"	"3535 Las Vegas Blvd"	Las Vegas	4774	American (New);Restaurants;Breakfast & Brunch	AMo5Nk8GAgm0OhQrnNm01A	I have been to all the HHGGs in town, and fina...
4	f4x1YBxkLrZg652xt2KR5g	"Hash House A Go Go"	"3535 Las Vegas Blvd"	Las Vegas	4774	American (New);Restaurants;Breakfast & Brunch	DNivSyKQ8OmYwCQWwmUQJw	Holy cow. Where do I start?\nFirst of all, I h...
...	...	...	...	...	...	...	...	...
7556	DkYS3arLOhA8si5uUEmHOw	"Earl of Sandwich"	"3667 Las Vegas Blvd S"	Las Vegas	4869	Caterers;Sandwiches;Restaurants;Food Delivery ...	_NDWz5_q-nSlzROSzQyliA	Ate too much for lunch and wanted just a snack...
7557	DkYS3arLOhA8si5uUEmHOw	"Earl of Sandwich"	"3667 Las Vegas Blvd S"	Las Vegas	4869	Caterers;Sandwiches;Restaurants;Food Delivery ...	s3ebYG1nZX89W5zt2YAifw	We had been to an Earl of Sandwich in Disneyla...
7558	DkYS3arLOhA8si5uUEmHOw	"Earl of Sandwich"	"3667 Las Vegas	Las Vegas	4869	Caterers;Sandwiches;Restaurants;Food Delivery ...	M_VPbaa-1fXXVtS3zUNRyw	This is one of my favorite spots in

Figure 6.1: Final Customer Reviews Dataset for multiple Restaurants

Final Customer Reviews dataset contains the ‘text’ attribute contains the customer review. This attribute is used for classification. Each text was split into single sentences and lower case, tokenization, stop words and removal of special characters were performed in order to prepare and preprocess the data.

6.3.1 Lowercase

Dataset contains the same word with lowercase or uppercase. This may deliver different output. Hence in order to preserve consistency this simple step is important in nlp tasks.

6.3.2 Stop words and Noise Removal

Stop words are known as the common words used in a language such as ‘a’, ‘the’, ’is’, ’are’ etc. Analysis tasks can be done efficiently by eliminating these words and focusing on important words. Sentences are also containing special characters, digits and other unnecessary text pieces. These things make noises for analysis tasks and should be removed.

6.3.3 Tokenization

Longer strings of texts should be split in to single sentences. If the text contains multiple sentences it should be split into one sentence.

6.4 Text/Review Classification using machine learning models

Text/review classification is an automated process of classification of texts into predefined categories. There are several machine learning techniques which can be used for text mining. This research work is implemented using machine learning techniques such as Naïve Bayes model, Support vector machine and deep learning methods with Convolution neural.

These models are to be trained and tested by annotated restaurant reviews of SemEval – 2014. Reviews are manually labeled with aspect categories. Dataset is split into training and testing datasets. These models will be used to classify the aspect category for the terms included in yelp review sentences. A classification models are used when terms cannot be assigned to broad aspect categories using wor2vec similarity functions. This situation occurs when the similarity values cannot reach the threshold value. Hence the aspect category for the review is detected using the trained machine learning model.

6.4.1 Naïve Bayes Classification method

This is a simple classifier based on probabilities of events of word occurrences in the corpus. Training datasets used for each categories (classes). Model can be trained to classify customer reviews for relevant categories. SemEval -2014 dataset contains reviews for several categories such as food, price, ambience and service. Some reviews cannot be categorized into a relevant restaurant aspect. Hence those are categorized as miscellaneous. Multinomial Naïve Bayes model is used in this research. It is a special instance of a naïve bays classifier which uses a multinomial distribution for each of the feature. The ML Naïve Bayes model is trained with restaurant reviews and aspect categories. Label Powerset Transformation is used to fulfill multi label classifications. Word frequencies are used as features for the model learning. Hence review sentences should be transformed in to vectors before feeding into the model.

Individual words are considered as independent of sentences and word frequencies are used for calculating probabilities. Techniques such as stop words, n-grams, Term-Frequency Inverse Document Frequency, Count vectors are used for vectorization.

The data is converted in to feature vectors and new features are generated using training data.

6.4.2 Text Classification using Support Vector Machines

It determines the best decision boundary between vectors that belong to a given category and vectors that do not belongs to it. Texts should be vectorized using techniques such as Term - Frequency Inverse Document Frequency and Count vectors. Stop words and n-grams are also used. Label Powerset Transformation is used since the problem is multi-label. These techniques are used for feature engineering. The data is converted in to feature vectors and new features are generated for further analysis. The model is trained using restaurant reviews and aspect categories.

Both Naïve Bayes and support vector machines are implemented with scikit-learn library.

6.4.3 Deep Learning Method

A deep learning model is trained with a suitable architecture and parameters for aspect category in this research. Convolution neural network is used for text classifications. Network architectures contain neurons organized in to multiple layers. Neurons in layers are connected with each other. The CNN model is designed with four layers and trained with training data of review texts and aspect categories. Vector representation of words is Bag-of-words. Dense class is used to define a fully connected layers where each neuron in the network receives input from all the neurons in the previous layers.

Maximum vocabulary size is created using word embedding and input shape is assigned 6000. Resulting data points are classified into multiple classes using nonlinear functions such as relu. The output layer consists of 4 neurons one for each class. Probabilities of class is returned using softmax activation function. Highest class is defined as the most suitable one. After processing the architecture, a configuration process is specified using an optimizer, loss function and accuracy metrics.

Model learning is finding a combination of model parameters that minimize a loss function for a given set of training data samples and their corresponding targets. Since this problem is multiclass, categorical cross entropy loss function is used.

Encoding is performed using Bag-of -words technique and tokenized words for each observation is used with finding the frequency of each token. Aspect categories are encoded using label encoder. The model is trained using fit function with tokenized reviews and encoded aspect categories. Keras library is used for model implementation classification.

6.5 Extracting terms of review sentences and generating dictionary with polarities for each term.

Techniques such as dependency parsing is used for aspect extraction. Spacy library is used for performing these tasks. Each review sentence is assigned to a parsing tree which identifies syntactical structures and relationships among words in the reviews. Nouns are considered as aspect terms and they are extracted.

Opinion lexicon with a collection of positive and negative words is used with a rule-based approach to detect the polarity of the opinion words associated with aspect terms. Finally, a dictionary with polarity score for each term is generated for further development.

6.6 Assigning aspect terms in to aspect categories and detecting the overall sentiment (positive/negativity)

Aspect categories (food, price, service, ambience) should be pre-defined. Extracted aspect terms should be assigned to broader aspect categories. Assigning process is done with word2vec model which is trained using Google's news dataset. This dataset is used for natural language processing tasks which contains bag-of-words and skip-gram architectures. Genism library is used for performing the training process of word2vec.

Word similarities are checked with functions with cosine similarity measures. It is the measure of distance between two words vectors in a vector space. The n_similarity function of word2vec is used for this calculation. If the assignment is not successful due to similarity value not reaching the threshold value, the trained machine learning model can be used for detecting aspect categories. Each aspect category finds the overall sentiments towards positivity and negativity. It is done by using all the sentiment scores associated with aspect terms towards an aspect category.

6.6.1 Finding most positive or most negative aspect terms of a category

They can be identified by the frequent terms which are being discussed by customers often and staff can take necessary actions for customer satisfaction. Dictionaries generated using spacy dependency parsers contains extracted aspect terms with polarity values. Terms with highest values are considered as frequent terms.

6.6.2 Visualizing Aspect categories of multiple restaurants for decision making

Final task is to generate a total value for the positivity of each aspect category. These categories are to be visualized by a simple graph in order to make comparisons among restaurants. Libraries such as Matplotlib is used for visualization.

6.7 Pseudocode for Extracting aspect terms and polarity detection

Simple pseudocode representation for extracting terms and detecting polarity values. Actual development process was a complex rule-based approach for extracting opinion words is used for this process.

```
Start
Input: Review sentence
        initialize the dictionary
        for each token in the sentence
            if opinion word is in pos
               Set sentiment = 1
             Else
              Set sentiment = -1
        Add sentiment to the dictionary
        If token = 'Noun'
           Add to the dictionary if not exist
        output: aspect terms dictionary with sentiment values
End
```

6.8 Pseudocode for Assignment of aspect terms in to aspect categories

```
Start
     Input: Extracted aspect terms dictionary
            Pre-defined aspect categories
     For each term in aspect terms dictionary
          Find aspect category
          If sentiment value > 0
                Increase positive sentiment value of aspect category
```

```
Else
            increase negative value of aspect category
   Output: aspect categories with aspect terms and sentiment values
End
```

This chapter presents the implementation details and technologies applied for the research. The next chapter will discuss about the results of the implemented system and evaluations.

Chapter 7

Results and Evaluations

7.1 Introduction

The topic in this research is to compare restaurant features based on past customer experience. Training dataset for machine learning models is from SemEval-2014. Restaurant reviews were collected from Yelp.com for the classification processes.

7.2 Measuring performance of classification models.

Machine learning models are generated for review classification. Models such as Multilingual Naïve Bayes, support vector machines, and deep learning convolution neural networks were created and the multi-label Naïve Bayes model received higher accuracy. Hence ML Naïve Bayes is used for training and classification.

Calculated accuracy of Naïve Bayes - 0.8664

	precision	recall	f1-score	support
0	0.86	0.26	0.41	121
1	0.69	0.84	0.76	289
2	0.82	0.71	0.76	301
3	0.94	0.42	0.59	80
4	0.80	0.62	0.70	145
micro avg	0.77	0.66	0.71	936
macro avg	0.82	0.57	0.64	936
weighted avg	0.79	0.66	0.69	936
samples avg	0.76	0.71	0.72	936

Table 7.1: Classification Report of Naïve Bayes model

Calculated accuracy of SVM - 0.8538

	precision	recall	f1-score	support
0	0.68	0.45	0.54	121
1	0.70	0.83	0.76	289
2	0.83	0.67	0.74	301
3	0.75	0.51	0.61	80
4	0.75	0.62	0.68	145
micro avg	0.75	0.67	0.71	936
macro avg	0.74	0.62	0.67	936
weighted avg	0.75	0.67	0.70	936
samples avg	0.76	0.71	0.72	936

Table 7.2: Classification Report of SVM model

Loss and classification accuracy of implemented Convolution neural Network.

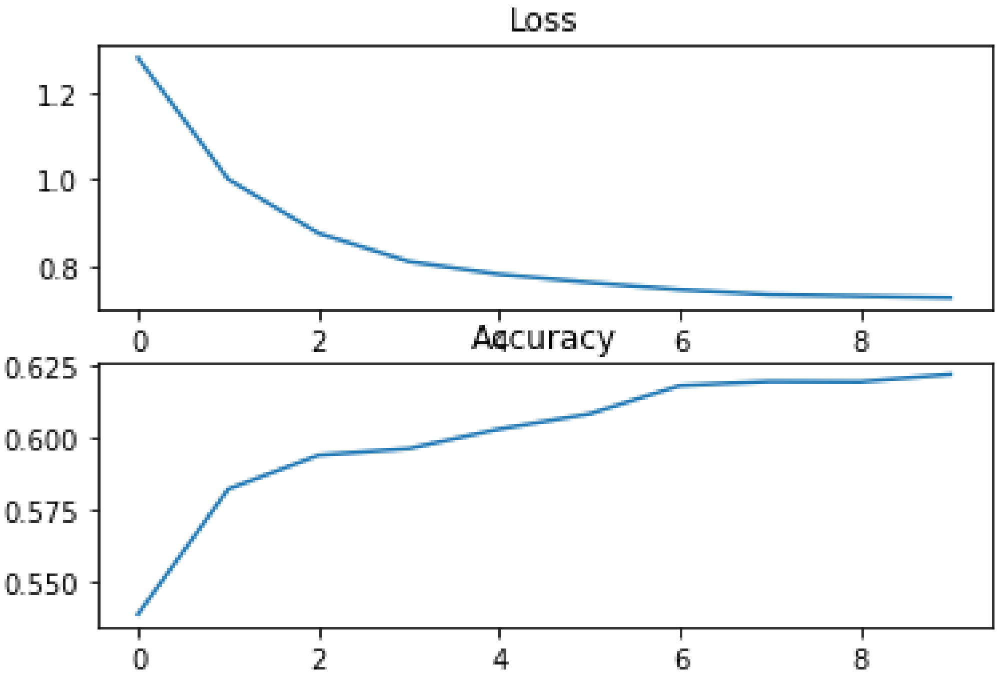
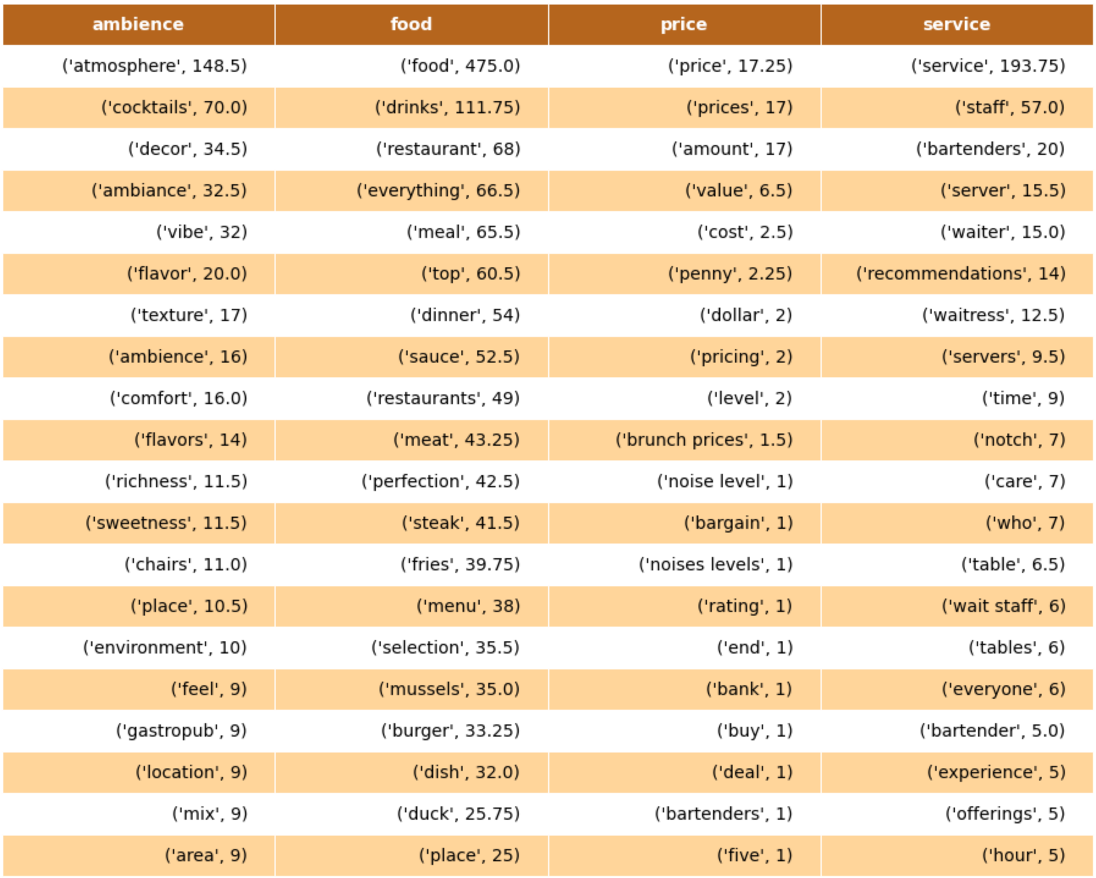


Figure 7.1: Loss and accuracy of Convolution Neural Network

7.2.1 Finding the most common and most positive aspects of a particular restaurant

This approach is useful for business staff. Frequent terms which are being praised by customers often or underestimated features can be identified by analyzing these terms. Staff can take necessary actions for customer satisfaction. Dictionaries generated using spacy dependency parsers contains terms with polarity values. Terms with highest values are considered as frequent terms. Following figure represents frequently used positive terms in Bacchanal Buffet restaurant.



ambience	food	price	service
('atmosphere', 148.5)	('food', 475.0)	('price', 17.25)	('service', 193.75)
('cocktails', 70.0)	('drinks', 111.75)	('prices', 17)	('staff', 57.0)
('decor', 34.5)	('restaurant', 68)	('amount', 17)	('bartenders', 20)
('ambiance', 32.5)	('everything', 66.5)	('value', 6.5)	('server', 15.5)
('vibe', 32)	('meal', 65.5)	('cost', 2.5)	('waiter', 15.0)
('flavor', 20.0)	('top', 60.5)	('penny', 2.25)	('recommendations', 14)
('texture', 17)	('dinner', 54)	('dollar', 2)	('waitress', 12.5)
('ambience', 16)	('sauce', 52.5)	('pricing', 2)	('servers', 9.5)
('comfort', 16.0)	('restaurants', 49)	('level', 2)	('time', 9)
('flavors', 14)	('meat', 43.25)	('brunch prices', 1.5)	('notch', 7)
('richness', 11.5)	('perfection', 42.5)	('noise level', 1)	('care', 7)
('sweetness', 11.5)	('steak', 41.5)	('bargain', 1)	('who', 7)
('chairs', 11.0)	('fries', 39.75)	('noises levels', 1)	('table', 6.5)
('place', 10.5)	('menu', 38)	('rating', 1)	('wait staff', 6)
('environment', 10)	('selection', 35.5)	('end', 1)	('tables', 6)
('feel', 9)	('mussels', 35.0)	('bank', 1)	('everyone', 6)
('gastropub', 9)	('burger', 33.25)	('buy', 1)	('bartender', 5.0)
('location', 9)	('dish', 32.0)	('deal', 1)	('experience', 5)
('mix', 9)	('duck', 25.75)	('bartenders', 1)	('offerings', 5)
('area', 9)	('place', 25)	('five', 1)	('hour', 5)

Figure 7.2: Most common and positive aspect terms

The following table shows sample values of totals of positive values and negative values for the Bacchanal Buffet restaurant.

pos_norm = positive/ (negative + positive)

neg_norm = negative/ (negative + positive)

	Negative	**Positive**	**Total**	**Neg_norm**	**Pos_norm**
ambience	123.750	426.500000	550.250000	0.224898	0.775102
food	1844.375	4172.890625	6017.265625	0.306514	0.693486
price	50.500	139.500000	190.000000	0.265789	0.734211
service	247.000	543.140625	790.140625	0.312603	0.687397

Table 7.3: Values of Positivity/Negativity for restaurant features.

7.3 Comparison for Restaurants found in Yelp.com

The comparison can be visualized in many ways using various types of graphs. Following figure shows an example for the comparison of four restaurants in Las Vegas. The graph helps users to make decisions over particular restaurant in a very short time period. Also, using this comprehensive information which is based on past customer experience and sentiments over restaurant features, business owners and their staff can make smarter decisions towards customer satisfaction without doing deep search. The graph is based on positive/negative values for each feature.

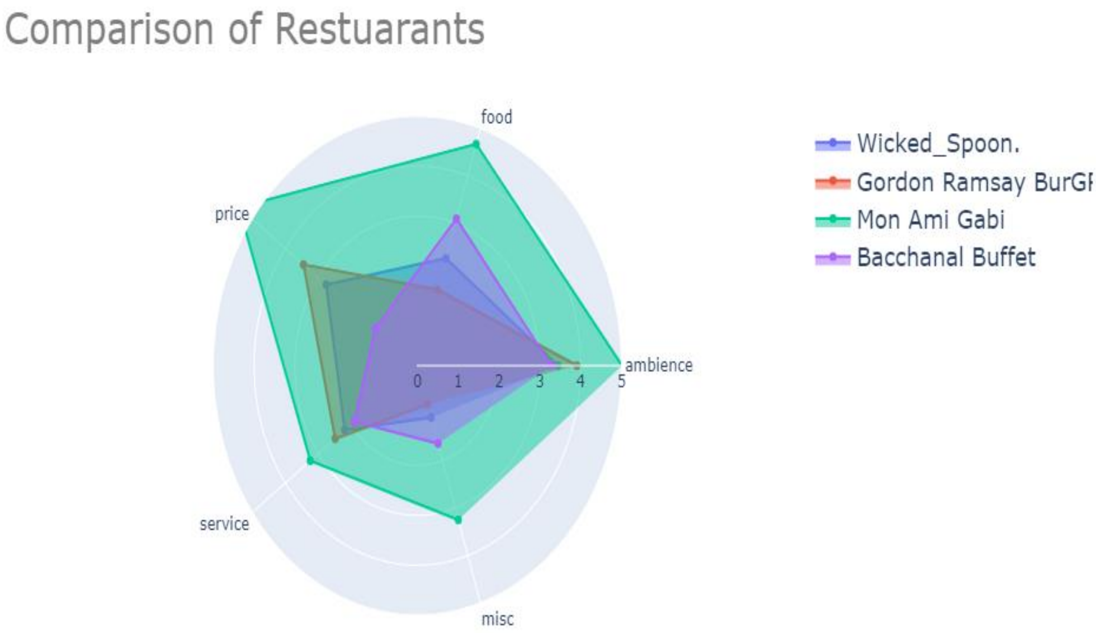


Figure 7.3: Visualization of Restaurant Feature Comparison

This chapter presented the results of the implemented system. The next chapter will present the discussion on the conclusion of the research.

Chapter 8
Conclusion

The aim of this research is to implement a reliable restaurant feature comparison system. Restaurants from Yelp.com are compared in this work. Natural language processing techniques which deal with human languages, are used for the development. The approach is based on aspect-based sentiment analysis. It is the word level analysis of a review sentence. Solution contains different development approaches. Machine learning approach together with rule-based approach is used. Hence the research contains a solution with hybrid technologies. Although three types of machine leaning models are used, Naïve Bayes model delivered better results over other models. Hence the research work is continued with that model. Support vector machine and convolution neural network model were the other models used for the training. SemEval – 2014 is used for training the models. Related work on this area was studied and many works were based on classification of restaurant features except comparison. Popular restaurant features were analyzed such as food, service, price and ambience. Development methodology had several steps such as model training and testing, term extraction from reviews and detecting the polarity of each term, assigning terms to aspect categories, finding overall sentiment towards broader categories and finally visualization of multiple restaurants comparing their features. Python based technologies used for implementations. Python is a common choice for natural language processing tasks present. Literature based on sentiment analysis proves it. Python tools such as spacy, genism, keras, scikit- learn are used. Both customers and restaurant staff can benefit from this work. Implementation results were presented in the previous chapter.

References

[1] Mubarok, M. (2017), *Aspect-based sentiment analysis to review products using Naïve Bayes,* in API Conference Proceedings

[2] Alqaryouti, O. (2019), *Aspect-based sentiment analysis using smar government review data,* ScienceDirect

[3] Nikolic, N. (2019)*, Aspect-based sentiment analysis of reviews in the domain of higher education*,Emerald Insight

[4] Jovelyn, C. (2018), *Text Minning customer reviews for aspect based resturant rating,* International Jounal of computer science and Information Technology

[5] Nurifan, F.(2019) *Aspect Based Sentiment Analysis for Restaurant Reviews Using Hybrid ELMoWikipedia and Hybrid Expanded Opinion Lexicon-SentiCircle*, Internationl Jounal of Intelligent Engineering and Systems

[6] Kok, S. D., *Review-Level Aspect-Based Sentiment Analysis Using an Ontology*.

[7] Kaynar, O. (2016), *Sentiment analysis with machine learinng techniques*, in International Artificial Intelligence and Data Processing Symposium

[8] Khan, A. (2016)*, Naïve Multi-label classification of YouTube comments using comparative opinion mining*, in Symposium on Data Mining Application

[9] Raybakov V , *Aspect-Based Sentiment Analysis of Russian Hotel Review*

[10] Hamdan H *, Supervised Methods for Aspect-Based Sentiment Analysis*

[11] Guha S. (2015) , *Aspect Based Sentiment Analysis in Reviews,* in Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)

[12] Luo Y. (2019), *Predicting the Helpfulness of Online Restaurant Reviews Using Different Machine Learning Algorithms: A Case Study of Yelp*, Sustainability

[13] Sindhu C. (2018), *Aspect Based Sentiment Analysis of Amazon Product Reviews,* International Jounal of Pure and Applied Mathematics

[14] Gojali S. (2016), *Aspect Based Sentiment Analysis for Review Rating Prediction*, in 2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)

[15] www.wikipedia.org

[16] Grabner D. (2012), *Classification of Customer Reviews based on sentiment analysis,* in 19th Conference on Information and Communication Technologies in Tourism (ENTER)

[17] Singla Z. (2017), *Sentiment Analysis of Customer Product Reviews*, in International Conference on Intelligent Computing and Control (I2C2)

[18] Asghar N. , *Sentiment Analysis of Customer Reviews based on Integration of Structured and Unstuctured Texts*

[19] Kritchenko S. (2014), *NRC-Canada-2014: Detecting Aspects and Sentiment,* in Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)

[20] Fang X. (2015), *Sentiment analysis using product review data*, Journal of Big Data,

[21] Bagheri A. (2013), *Care more about customers: Unsupervised domain-independent aspect detection for sentiment analysis of customer reviews*, Elsevier

[22] Jurgen, *Aspect-Oriented Sentiment Analysis of Customer Reviews using Distance Supervision Techniques*

[23] Ruder S. (2016), *A Hierarchical Model of Reviews for Aspect-based Sentiment Analysis*

[24] Toh Z. (2016), *NLANGP at SemEval-2016 Task 5: Improving Aspect Based Sentiment,* in Proceedings of SemEval-2016

[25] Baharudin B. (2011), *Sentiment Classification from Online Customer Reviews Using Lexical Contextual Sentence Structure*, in International Conference on Software Engineering and Computer Systems