National College of Ireland

Recommender system for food in a restaurant based on Natural Language Processing and Machine Learning

MSc Research Project
Data Analytics

Kedar Ratnaparkhi
x17111013

School of Computing
National College of Ireland

Supervisor: Dympna O'Sullivan

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Kedar Ratnaparkhi
Student ID:	x17111013
Programme:	Data Analytics
Year:	2018
Module:	MSc Research Project
Lecturer:	Dympna O'Sullivan
Submission Due Date:	13th August 2018
Project Title:	Recommender system for food in a restaurant based on Natural Language Processing and Machine Learning
Word Count:	XXX

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 encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	13th August 2018

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. 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):	

Recommender system for food in a restaurant based on Natural Language Processing and Machine Learning

Kedar Ratnaparkhi
x17111013
MSc Research Project in Data Analytics

13th August 2018

Abstract

Millions of customers post online reviews in the form of their own experience at restaurants. Some are positive while some are negative. Usually an overview of all the reviews is provided on the respective restaurant page. But this approach is hardly accurate or efficient. This research analyzes user reviews in restaurant domain, and then consolidates the information recommending the best dishes served to a customer at a restaurant. The system is developed using modern NLP techniques such as sentiment lexicon, sentiment scores, POS tagging to generate useful features and classify the information using Machine Learning classification algorithms such as KNN, Random Forest and SVM. The system achieved more than 93% accuracy across various experiments, with Random Forest performing best for the given dataset and SVM giving the best performance with a cross-validated dataset.

Keywords:*Text Mining, Natural Language Processing, Customer Experience, Machine Learning, Restaurant Review, Social Network, Lexicon, Sentiment Analysis*

1 Introduction

Many online web portals allow customers to share their experience and rate a given restaurant based on it. The amount of reviews generated is huge and it is an enormous task to analyze them and generate some useful information that a new customer can utilize.

In previous implementations, different approaches have taken place in order to analyze restaurant reviews by customers by using Natural Language Processing and Machine Learning techniques. Few of the published studies are reviewed here, such as by Zhang et al. (2015), which summarizes user descriptions for various dishes by means of LDA topic modelling, Chinsha and Joseph (2015) built a system which finds sentiments related to various aspects of a restaurant such as the food, service, price, ambience, by making use of the opinionated words used in each review sentence. Dayan et al. (2015) built a clustering based system which clusters restaurant with similar food genres and finds unique features in each restaurant genre. These researches are reviewed in detail in later section.

The problem with these approaches is that they try to either summarize the overall aspects of a given restaurant or recommend similar restaurant based on a user's previous visits. But this research takes a novel approach in building a system which recommends a user with a set of dishes which have been classified into three broad categories such as Good, Average and Bad based on previous reviews.

The aim of this research will be to answer the following question:

"Using Machine Learning and Natural Language Processing techniques, how can we distinguish the best dishes served at a restaurant from the bad ones?"

2 Related Work

Since this research focuses on improving the customer experience at restaurants by recommending the best dishes served at a given restaurant by analysing text reviews, previous studies from the field of text analytics, Natural Language Processing and Machine learning algorithms are studied which are prominently based in hospitality industry and improving customer experience.

This section is further divided into subsections based on the broad categories that are reviewed, such as 2.1 Restaurant Rating, 2.2 Aspect based Sentiment Analysis and 2.3 Classification based on Ratings, in order to gain knowledge and answer the research question under discussion.

2.1 Restaurant Rating

These days, social mediums are used for various purposes, from sharing day-to-day incidents, publishing news articles, blog posts to sharing opinions on various topics. One such use of social medium is sharing opinions about customer experiences at public places such as Restaurants, cafes, etc. Many researchers, with the use of modern text analytical technologies, make use of this huge data for the purposes of building recommendation engines, rating existing businesses, which is done by processing the textual reviews posted by previous customers, which could be useful for the future customer.

A few studies such as Chinsha and Joseph (2015) and Panchendrarajan et al. (2017) focuses on rating the restaurants based on various aspects of the restaurant based on user reviews. Although a similar approach is taken in rating restaurants, the way aspects are identified is different in both the studies. In Chinsha and Joseph (2015), the aspects are considered as the words which are tagged with POS tags NN/NNP/NNS by the Part of Speech Tagger using *The Stanford Parser: A statistical parser* ((Accessed July 28, 2018), and then aspect level sentiment polarity is calculated. Whereas, Panchendrarajan et al. (2017) uses human participants to manually annotate a set of reviews with aspects, sub-aspects in a pre defined hierarchy. The system also uses a pre-defined dictionary of words that is used to find the food item names, present in the review text. A similar idea has been taken when implementing this research, to identify the text snippets containing food reviews. Once the food dishes are found in the review text, they are then categorized into various clusters, according to their similarity, by the means of Single Pass Partitioning Method. Once all the aspects, sub-aspects, food categories are identified, sentiment scores are generated for each hierarchy level for that restaurant. Although presents an impressive implementation, the system relies heavily on the manual annotation of the training set of reviews, which could be substituted by an automated method.

In the study by Zhang et al. (2015) a short snippet/summarization for each food dish is provided by the system, which makes use of Bilateral Topic Analysis model. The proposed technique also generates a score for the overall review based on the descriptive words present in the review. The system then outputs a list of snippets describing a particular dish, for the user to read along with an overall score. By making use of LDA and Topic modeling, the system splits the review text into snippets and filter only those snippets containing the food dish names. A similar approach is implemented in this research, for filtering the N-grams containing food dish names.

Building recommendation systems based on the users' previous behaviour using various techniques is also very popular amongst researchers such as Dayan et al. (2015) and Trevisiol et al. (2014), which recommends restaurants and food menus based on previous user experiences, respectively . A two phase iteration process based on TF-IDF weighting scheme and Affinity Propagation Algorithm by Dayan et al. (2015), clusters restaurant serving dishes with similar food genres. Although, this is not a novel idea. What makes it a novel approach, is during the second iteration of the above process, the system extracts unique features amongst those clusters, recommending the most unique restaurants in any given category to users. The study by Trevisiol et al. (2014) recommends food menus, instead of food dishes or restaurants, which are ordered together by the previous customers by using Fuzzy Apriori Algorithm. In order to find all the food items from the given reviews, various open source dictionaries containing food dish names are used, which served as a motivation for using a similar food dictionary/ontology in this research. Another use of Fuzzy representation was done by Sauper and Barzilay (2013), where a probabilistic model is used to detect aspects related to a specific restaurant category and the sentiments associated with the identified aspects from the reviews text. Using Fuzzy representations, the aspects and the sentiment polarity related to the aspects are extracted together, rather than splitting the task in two separate steps.

2.2 Aspect based Sentiment Analysis

Sentiment Analysis has been applied in many previous studies which includes analysing user sentiments in terms of online reviews, opinions posted online towards certain product/service. In the studies such as Hossain et al. (2017) and Chinsha and Joseph (2015), the primary motivation was to analyse the overall user opinion expressed for a given restaurant and classifying the reviews as either positive and negative. Although, this is a very efficient approach towards sentiment analysis, it is also a very crude method, which only gives an aggregated result. In order to get a more granular representation, a more efficient approach is to perform aspect/feature based sentiment analysis.

The main problem when dealing with aspect based sentiment analysis is to first of all find the aspects from the given review text and then extract the sentiments associated with these aspects as accurately as possible.

Hence, aspect based sentiment analysis is being studied by researchers since the past few years. Many a times the terms, Aspects and Features are used interchangeably, but they refer to the same concept. Aspects could be considered as a particular property or feature of a given service/product, towards which users can express their opinions. Aspects can be identified manually from a given text by performing various Natural Language Processing techniques as demonstrated by Akhtar et al. (2017), such as POS tagging, Head Word identification, frequency of occurrence of certain words, Stop word identification, Lemmatization are used to identify aspects. Such a comprehensive tech-

nique can be used for manually identifying aspects in any domain, but is very time consuming and requires a lot of resources for aspect term identification, even before the actual task of sentiment generation starts. A semi-supervised approach used by Garca-Pablos et al. (2018) takes one domain dependent word for each aspect of a particular product/service that needs to be analyzed, along with a word with positive and negative polarity. These seed words are then passed to a system called Word2Vec(Mikolov et al.; 2013), which makes use of a technique called Word Embedding, which finds words with semantically similar meanings. Using this technique, it builds itself a vectorized list of aspect words similar meanings to the seeded word.

Another approach to tackle this problem of aspect identification is to use a pre-built dictionary/ontology as an input to the system, which would contain a list of domain specific features/aspects. The use of such an approach is taken by Pealver-Martinez et al. (2014), in order to identify aspects from the movie domain and analyse user sentiments towards these aspects. An ontology containing various features related to movies such as actor names, movie names, genres, etc is used in this particular study. The use of such approach is very useful since the system can be used for performing similar tasks for any domain, just by changing the input dictionary/ontology which is domain dependent. From the review of above studies, it is learned that various approaches can be taken for feature/aspect extraction, either Manual generation by processing the textual data or by using an ontology/dictionary based methodology. For the purpose of this study, the decision of using a dictionary based approach is taken which contains a list of food items. The same will be discussed in upcoming sections.

Once the aspects/features are identified from the review text, the next sub-task is to analyse the sentiments associated with these aspects. In the researches performed by Panchendrarajan et al. (2017), Chinsha and Joseph (2015), once the aspects such as food, ambience, service, etc are identified, the sentiments associated with these aspects are extracted and a sentiment score is generated. In order to generate a sentiment score, various lexical resources are used, which are used in different contexts, such as sentence level(Appel et al.; 2016), document level(Sharma and Dey; 2012), N-gram level(Kang et al.; 2012). Guerini et al. (2013) performed various calculations using the scores from such lexicons such as aggregation, mean, highest occurring words, etc, in order to generate sentiment scores for a given aspect.

All the above studies generates various aspects related to restaurant domain such as, food, ambience, price, service, etc and then perform a sentiment analysis workflow on the extracted data. But in this research, a similar approach is taken, but implement with a different ideology, for extracting food dishes as aspects and then generate features related to sentiment scores related to those aspects.

2.3 Classification based on Ratings

In order to build a model that correctly classifies the occurrence of future cases based on the past ones, Machine Learning algorithms are used in order to build supervised or unsupervised systems. Sentiment classification is one of the most popular application implemented using classification algorithms. Many studies such as Bhattacharjee and Petzold (2017), Hossain et al. (2017) and Tripathy et al. (2016) use different classification algorithms to perform sentiment classification into two classes i.e. Negative and Positive. SVM, KNN, Random Forest and Naive Bayes are one of the most popular and easy to interpret classification algorithms available today for academic or commercial use.

There are two approaches that can be taken, in order to perform sentiment classification of user reviews:

- Generating Vector representation of the given review sentences by using TF-IDF such as used in Mouthami et al. (2013), Word2Vec such as used in Bhattacharjee and Petzold (2017), and passing them as feature sets in order to train classification models such as Logistic Regression(Bhattacharjee and Petzold; 2017), SVM(Mouthami et al.; 2013)

- The other approach is to generate features manually, from the available free text data using various Natural Language Processing techniques as performed by Akhtar et al. (2017), where more than 18 features are generated using techniques such as Lemmatization, POS tagging, Word Length, etc. In the work of Patra et al. (2015) as well, various features are generated such as Positive/Negative polarity of words, POS tags, Word occurrence frequency, Number of sentences in a given review. In Chinsha and Joseph (2015), use of SentiWordNet(Esuli and Sebastiani; 2006) lexical resource is done, using which positive and negative sentiment polarities for various aspects are generated as features to be passed to the classification model.

Once these feature sets are generated, the classification algorithms are trained for sentiment classification, such as SVM used in Zhu et al. (2016), Hossain et al. (2017) and Mouthami et al. (2013), Random Forest, which uses an ensemble approach for prediction purposes, is used in Zhu et al. (2016) and Wan and Gao (2016). K-nearest Neighbour is also said to give good classification accuracy of around 74% in Hossain et al. (2017) for sentiment classification.

By reviewing the above studies, various feature sets are generated from the reviews text data and are classified using SVM, Random Forest and KNN as they have proved to give good results.

3 Methodology

In implementing this research, KDD Methodology as proposed by Fayyad et al. (1996) is followed, which consists of a set of defined steps that can be taken in order to implement any research project which aims at generating knowledge from a given dataset. The same can be seen from Figure 3. The rest of this section will explain the architecture of this study in brief overview which follows KDD life cycle:

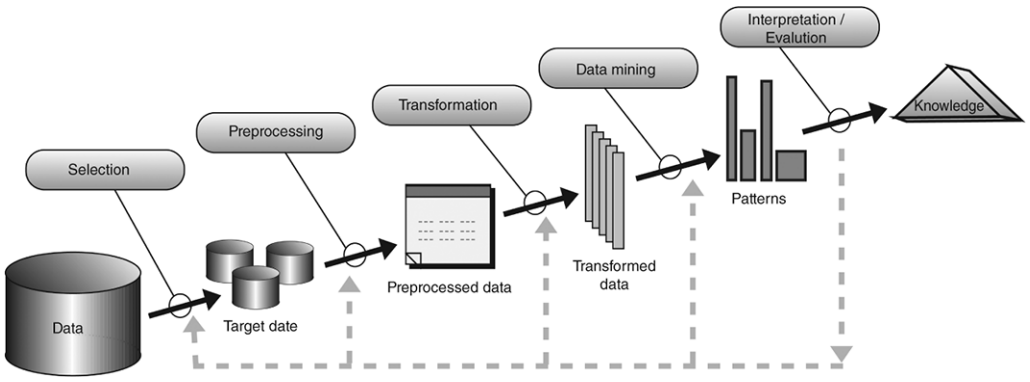

Figure 1: KDD Methodology (Fayyad et al.; 1996)

3.1 Data Selection and Collection

Many popular datasets from the restaurant domain are used in previous research which primarily focuses on sentiment analysis. A few of the datasets available are already annotated with some useful metadata such as sentiment score, sentiment polarity, aspects contained in the reviews, which promotes the researchers with some extra information to work on. SemEval 2016 is one such examples which has been used in Garca-Pablos et al. (2018). The most popular datasets available online for public use is (*Tripadvisor*; (Accessed July 28, 2018), which provides user reviews on their website, which was used in Chinsha and Joseph (2015). Although the volume of reviews from *Tripadvisor* ((Accessed July 28, 2018) is quite large, acquiring the data requires performing web scraping of the review pages manually. Hence, the use of Tripadvisor was discarded for the purpose of this study, considering the time frame available.

Apart from the above data sources, another open source dataset available from *Yelp* ((Accessed July 28, 2018) is also used in various studies, which focuses on the domain of sentiment analysis, such as Bhattacharjee and Petzold (2017), Zhu et al. (2016). The same has been used in this study as well.

3.2 Data Pre-Processing

The dataset from *Yelp* ((Accessed July 28, 2018) contains more than a million rows which contains reviews for not only restaurants, but also reviews from other businesses, such as barber shops, groceries, etc. For the purpose of this study, the reviews are filtered and only the reviews from restaurant category are extracted. From this dataset, the reviews for 15 restaurants with the maximum number of reviews are selected for this study, which is around 70K.

Once the necessary data is filtered, the reviews text are then processed by converting all the text into lower case, to keep the data consistent. Also, stop words such as the, is, at, etc are removed from the review text, since these words contain no extra information for the analysis. Another benefit of removing stop words is that it reduces the amount of data that is to be processed, significantly, improving system efficiency.

3.3 Data Transformation

Data is extracted as described in Table 2 and transformed into a format that can be fed into various machine learning classification algorithms. The study from Akhtar et al. (2017) proved to be of major inspiration while generating many of the features generated during this step. Machine learning algorithms understand the language of 0s and 1s and cannot interpret plain text data written natural language. Hence, the given textual data is transformed into a set of numerical features, that could be understood by computers using following techniques:

- **Seed Information:**
  Most of the studies reviewed so far which performs the task of aspect/feature based sentiment analysis as part of their project, makes use of some kind of dictionary/ontology as an input to the system, using which the aspects/sentiments from the review text are identified. Studies such as Zhang et al. (2015) and Panchendrarajan et al. (2017) made use of a food dictionary/ontology to identify food dishes

from the given review text. Bhattacharjee and Petzold (2017) used the restaurant descriptions like parking, food, ambience, location, etc. as aspects which are provided to users on Yelp.com. Hence, in this research also, a similar approach, although in a different way is taken, where a list of food words is passed to the system, to find the sections of the review mentioning food.

- **Sentiment score:**
  In order to generate a sentiment score of the opinionated words, a traditional approach can be taken to generate positive or negative polarity score of the opinionated word, as done in . But in order to get a more accurate and precise sentiment score of the opinionated words, few studies: Staiano and Guerini (2014), Patra et al. (2015), Pealver-Martinez et al. (2014) have used some of the open source sentiment lexicons such as Sentiwordnet(Esuli and Sebastiani; 2006), MPQA Lexicon(Wiebe and Mihalcea; 2006), Bing Liu(Ding et al.; 2008).

  Hence, in this research, the use of sentiment lexicon, SentiWordNet(Esuli and Sebastiani; 2006) is done to generate sentiment scores in a given review.

Once all these steps have been performed, the data is transformed into a machine understandable format as shown in below table.

dish_name	rest_name	stars	sent_score_pos	sent_score_neg	overall_score_pos	overall_score_neg	food_freq	class
appetizer	Mon_Ami_Gabi	4	0.619047619	0.005952381	0.330315518	0.066974692	1	3
beef	Mon_Ami_Gabi	3.666666667	0.28546627	0.051773313	0.198193122	0.075017982	6	2
crab	Mon_Ami_Gabi	5	0.078125	0.541666667	0.199107143	0.319642857	2	3
mimosa	Mon_Ami_Gabi	4	0.221875	0.075	0.188229167	0.073303571	2	3
omelet	Mon_Ami_Gabi	4.333333333	0.159722222	0.163194444	0.196627784	0.086916763	3	3

Table 1: Transformed Data

3.4 Data Mining

In order to correctly classify the instances based on previously labeled cases, Supervised machine learning approach is taken in this research. The most popular and proven to provide accurate results in multi class classification problems, by Zhu et al. (2016), Mouthami et al. (2013) and Wan and Gao (2016) are Random Forest, K-Nearest Neighbour and SVM. The primary focus of this study is to classify the dishes served at a particular restaurant, but due to scarcity of data from a single restaurant, various other experiments are carried out, where sampling using Cross Validation and combined reviews data from various restaurant are used to build a more powerful model.

3.5 Interpretation and Evaluation

Once the classification models are trained with the testing data, various evaluation metric, such as Accuracy, Specificity, Sensitivity, etc will be measured to evaluate the performance of the classification models. Also, since multi class classification is being performed in this study, confusion matrices of the results will also be checked to interpret the performance of the model for each class.

4 Implementation

In order to build the system which recommends the best dishes to a customer, the system needs to analyse the given set of reviews and then classify the food dishes into three pre decided classes. The following steps are carried in order to perform the classification:

1. Data Pre-Processing
2. Feature Generation
3. Model Classification

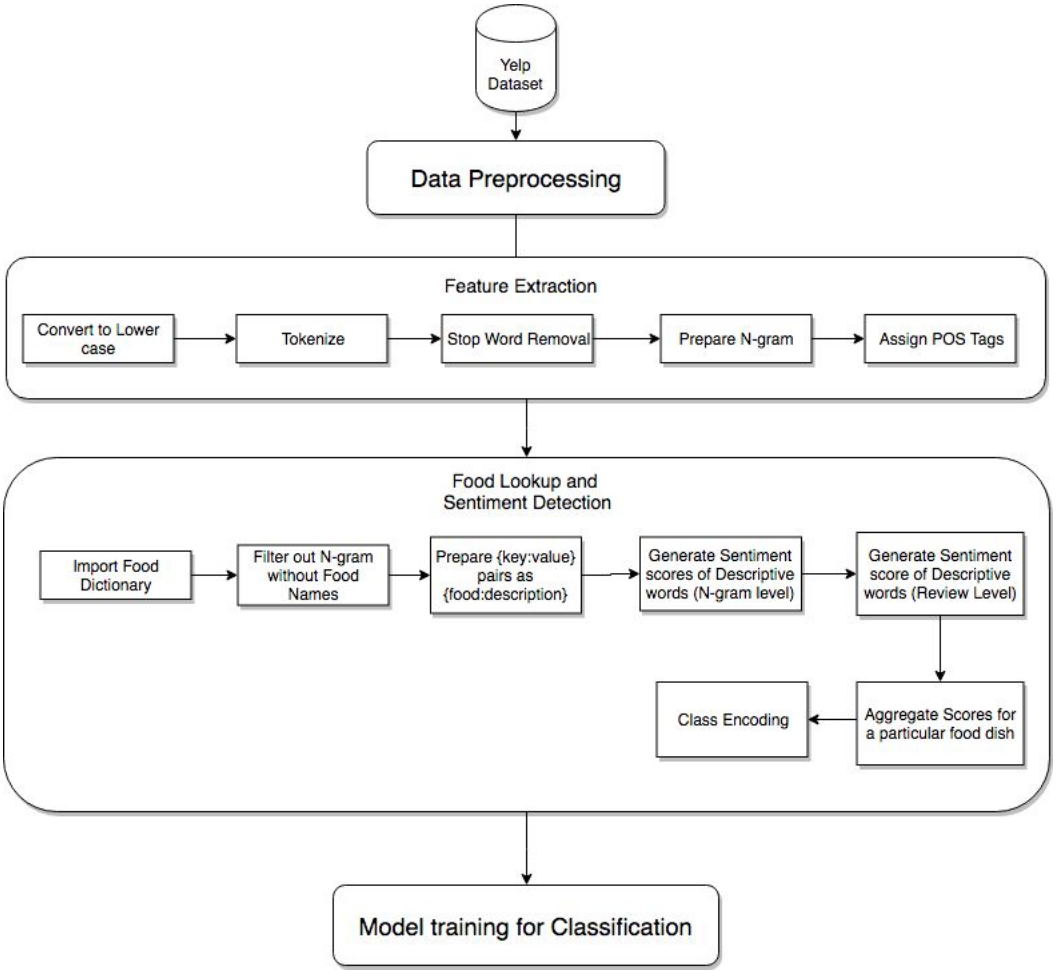

Figure 2: Implementation Steps Overview

4.1 Data Pre-Processing

The restaurant reviews data is downloaded from *Yelp* ((Accessed July 28, 2018) in a well structured tabular SQL format. Since the dataset downloaded was divided into various tables, for this research only the table containing reviews was used and all the other tables were excluded from the process. The dataset contained user reviews for not only restaurants, but also other businesses such as Grocery stores, barber shops, etc. Hence only those reviews which were generated against businesses which fall under "Restaurant" category were filtered and selected for analysis. Any columns which could

be used to identify the person who posted the review are excluded in order to avoid any ethical issues. Then, reviews for a single restaurant is selected at a time, and extracted into a CSV file in following format.

Restaurant Name	Text	Stars
Charr An American Burger Bar	Used them for delivery and I was highly upset with the overall experience. Looked at their menu to order crab cakes and...	5
Charr An American Burger Bar	...eat both times so I'm not sure what the others are complaining about. They,have a few decent craft beers...	3
Charr An American Burger Bar	The spicyness of the wings does build up and make it almost unbearably hot (to me) but the chili sauce they put on them is delicious	3

Table 2: Preliminary Data

This data is then further cleaned to remove special characters, stop words using basic python libraries, reducing the number of words to be processed further. Since we are analysing the dishes served at a particular restaurant, the data for each individual restaurant is processed one at a time.

4.2 Feature Generation

Majority of the features that will be used further to classify the dishes into predefined classes i.e. Good, Average and Bad have been generated during this process. Many of the studies reviewed above performs the task of sentiment analysis on a data which already contained annotated aspects and sentiment polarity for each review. In this research, all this was done from scratch, without any need for manual human intervention. The following steps were performed on the user reviews extracted from *Yelp* ((Accessed July 28, 2018) dataset and a new dataset is created containing all these generated features:

- **Prepare N-gram and Tokenization**
  Once the reviews are cleaned and prepared in a usable format, they are then split up into N-grams, which is a set of n number of adjacent words. In the study performed by Ashok et al. (2016), the review text is split into sentences, and analysed on sentence level, but in this research, the polarity is calculated per N-gram, since it limits to the number of words that needs to be analysed when generating sentiment score for that N-gram, which saves computational resources. The assumption behind this technique is that in a natural language sentence, the descriptive words are used just before or after a noun in order to describe the same. In case of user reviews, many different values of N-grams are tried, and the use of 3-grams is found to give relevant results. Once a set of N-grams are prepared for a given review, the N-grams are then split up into single words so that they could be analysed individually.

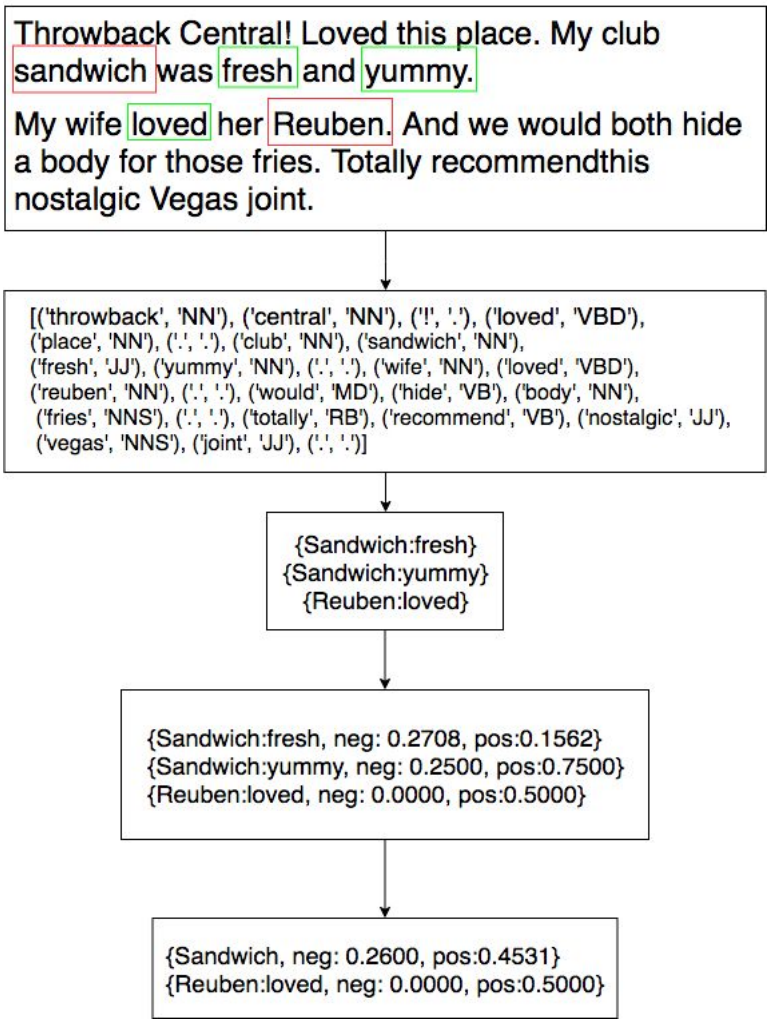

Figure 3: Feature Generation Steps

- **POS Tagging**
  Each N-gram is then passed to the Part-Of-Speech tagger, which appends each word, its part of speech depending on English language rules. This step is performed by using Stanford POS tagger, which is available in the Python NLTK package. After this step, we have more information generated regarding each word, which will help us analysing the given sentence in granular detail. In order to find the sentiments mentioned towards a given dish, only those N-grams are filtered which contain any adjective words i.e. Words tagged with the POS tag:**JJ**.Refer Figure 4

Table 2
The Penn Treebank POS tagset.

1. CC	Coordinating conjunction	25. TO	*to*
2. CD	Cardinal number	26. UH	Interjection
3. DT	Determiner	27. VB	Verb, base form
4. EX	Existential *there*	28. VBD	Verb, past tense
5. FW	Foreign word	29. VBG	Verb, gerund/present participle
6. IN	Preposition/subordinating conjunction	30. VBN	Verb, past participle
7. JJ	Adjective	31. VBP	Verb, non-3rd ps. sing. present
8. JJR	Adjective, comparative	32. VBZ	Verb, 3rd ps. sing. present
9. JJS	Adjective, superlative	33. WDT	*wh*-determiner
10. LS	List item marker	34. WP	*wh*-pronoun
11. MD	Modal	35. WP$	Possessive *wh*-pronoun
12. NN	Noun, singular or mass	36. WRB	*wh*-adverb
13. NNS	Noun, plural	37. #	Pound sign
14. NNP	Proper noun, singular	38. $	Dollar sign
15. NNPS	Proper noun, plural	39. .	Sentence-final punctuation
16. PDT	Predeterminer	40. ,	Comma
17. POS	Possessive ending	41. :	Colon, semi-colon
18. PRP	Personal pronoun	42. (	Left bracket character
19. PP$	Possessive pronoun	43.)	Right bracket character
20. RB	Adverb	44. "	Straight double quote
21. RBR	Adverb, comparative	45. `	Left open single quote
22. RBS	Adverb, superlative	46. ``	Left open double quote
23. RP	Particle	47. '	Right close single quote
24. SYM	Symbol (mathematical or scientific)	48. ''	Right close double quote

Figure 4: List of Parts Of Speech tags Marcus et al. (1993) which are used in Stanford POS tagger to assign to various Parts of Speech in a given sentence.

- **Food Look-up and Detect Sentiments**
  In order to find the food dishes included in a given review text, there are two options available. Either compile a list of food dishes served at each individual restaurant and then search for those dish names in the reviews of that particular restaurant, or use a predefined generic dictionary. The second option was chosen since it is much more efficient, less time consuming and can be used in a generic way for all the restaurants. In this study, Food dictionary from *WordNet* ((Accessed July 28, 2018) is used, which contains, not only food dish names, but also ingredient names as well, such as ginger, milk, etc. The ingredients were filtered out by understanding the schema of the dictionary, to get more relevant information. N-grams which does not contain any dish names are then filtered out, since they are not relevant for this research.

  So, at the end of this step, only those N-grams are left which contain any food dish name and nearby adjectives used in that sentence. Once this is done, a set of {key:value} pairs are generated for a given review, where key is the food dish name whereas the value is the adjective word found in that N-gram. Refer Figure 4.2. If in a given N-gram, there are multiple adjectives, those many number of pairs will be generated. At the end of this step, the system has generated a set of {key:value} pairs, which may contain some duplicates. These are filtered out as well to get a distinct set of {key:value} pairs. A similar approach is taken by Bhattacharjee and Petzold (2017) as well.

- **Sentiment Score Generation and Food Frequency**
  Once all the {food dish:sentiment word} pairs are generated, a sentiment score of each adjective word is then generated by using an open source Lexicon resource called Sentiwordnet(Esuli and Sebastiani; 2006) is used which provides a positive, negative score for most of the opinionated and subjective word which is an adjective, present in the English language dictionary. Sentiword has also been used for identifying sentiment polarity in a few researches like Patra et al. (2015), Ashok et al. (2016), Akhtar et al. (2017). For this research, an aggregated positive and negative sentiment scores of dish level sentiment words as well as entire review level aggregated positive and negative sentiment scores are generated.

While processing the data for the above feature generation, the frequency of that particular food dish mentions throughout the reviews is also calculated. The idea behind measuring this information is that if a dish has been mentioned more number of times, the dish might be more popular among customers or could be the worst dish served at a restaurant and the same has been expressed by many customers and may be a useful parameter to classify the food dish. A similar approach is taken in the work done by Patra et al. (2015) while generating one of the features used for sentiment classification.

- **Class Encoding**
  In order to label the dishes into respective classes, the rating stars given by the users in the original review are used. Since, the dishes are to be classified into three classes i.e. GOOD, AVERAGE and BAD. A similar approach was taken by Bhattacharjee and Petzold (2017) while classifying restaurant reviews into various classes. The reviews with star rating 1 is encoded as BAD, reviews with star ratings 2 and 3 are encoded as AVERAGE and reviews with star ratings 4 and 5 are encoded as GOOD.

- **Training Machine Learning Algorithm**
  Once the data has been transformed in the required format, various classification algorithms such as K-Nearest Neighbour with different values of K, Support Vector Machine (SVM) and Random forest is then trained on the resultant dataset and the most important evaluation metrics are compared with each others, to find the best suitable model for predicting the classification of food dishes into various classes. The results of the experiments will be discussed in the next section, with necessary details.

5 Evaluation

The aim of this research is to find the best food dishes served at a given restaurant, for which the textual reviews of previous customers are analyzed and various sentiment scores are computed. Once all the features are generated, supervised machine learning classification approach is taken. In order to serve the purpose of this research, classification models such as K-Nearest Neighbour, Support Vector Machine (SVM) and Random forest are trained with the generated data and the most important evaluation metrics are compared to find the model which predicts the dishes most accurately.

Two different experiments are carried out, where the reviews of individual restaurants are analysed and also by combining the data from all the hotels under analysis. The results are then compared with each other to find the best suited approach for this kind of analysis.

5.1 Experiment 1: Single Restaurant Analysis

In order to find the best dishes served at a given restaurant, the primary approach to the experiment is to train the classification models with data from individual restaurants. The same has been performed with a few restaurants selected randomly. But due to the scarcity of the number of reviews, the dataset could not be divided properly into training and testing subsets so that both the subsets would have dishes with all the

classes. Hence, very few restaurants could be analysed with this approach. The results of this experiment has been documented below for one of the restaurants selected randomly.

Here, from the results it can be seen that due to class imbalance in the datasets of individual restaurants, SVM predicted the cases with very low accuracy levels, as was also observed by Patra et al. (2015).

	Reference	
Prediction	Average	Good
Average	0	0
Good	4	11

Figure 5: SVM confusion matrix

Similar results were seen with K-NN as well, with varying values for K. Random forest proved to be the most accurate model used for predicting the cases even with very less number of cases to train with. An accuracy of 93.33% is achieved for one of the restaurants. As can be seen, this particular restaurant did not have any dishes with rating "Bad", hence only the dishes with "Average" and "Good" rating were classified. Hence, this approach is less than optimal in order to perform the required analysis.

Model	Accuracy	Sensitivity	Specificity	Kappa
K-NN with k=7	80	0.25	1	0.3284
Random Forest	93.33	0.75	1	0.8148
SVM	73.33	0	1	0

Table 3: Evaluation with Individual Restaurant

5.2 Experiment 2: Analysis on data from top 15 Restaurants

For the second experiment, the processed data from 15 restaurants with the most number of reviews are combined, and is used to train the classification models. In the first experiment, where the analysis was being done separately for each restaurant, the dataset size was very small to train the model efficiently. Hence, another experiment is carried out with the larger dataset, which is created by combining the data of the top 15 restaurants to train the models under consideration.

Since, the factors other than "Restaurant Name" are used for training the models, the final result of the model is being predict without any dependency with the restaurant name. To keep things consistent and easy for comparison, the same models are used for the second experiment as well. It is found that SVM works best when the classes are well balanced in the dataset under analysis. In this dataset, the number of cases where the dishes are in class 1 i.e. "Bad" are less than 30, SVM failed to classify any of them in the correct category, whereas the other classes are predicted correctly with varying accuracy. Similar behaviour was observed by Patra et al. (2015).

A slightly better prediction is observed when K-NN is used with a k-value of 7.

	Reference		
Prediction	Bad	Average	Good
Bad	5	0	0
Average	1	48	4
Good	0	16	181

Figure 6: K-NN confusion matrix with k=7

But the best performing model found to be is Random Forest, which predicts the correct classes with an accuracy of 93.33%. Below is the confusion matrix where the model correctly predicts the dependent variable even when there are very few cases for a given class.

	Reference		
Prediction	Bad	Average	Good
Bad	6	0	0
Average	0	56	8
Good	0	8	177

Figure 7: Random Forest confusion matrix

Below are the results with the most important evaluation metrics which are compared with the three classification models used for classification purposes.

Model	Accuracy	Sensitivity	Specificity	Kappa
K-NN with k=24	91.37	Class Good:0.98 Class Average: 0.78 Class Bad: 0.00	Class Good: 0.80 Class Average: 0.95 Class Bad: 1.00	0.7744
K-NN with k=2	90.98	Class Good: 0.94 Class Average: 0.81 Class Bad: 0.83	Class Good: 0.82 Class Average: 0.94 Class Bad: 1.00	0.7776
K-NN with k=7	91.76	Class Good: 0.97 Class Average: 0.75 Class Bad: 0.83	Class Good: 0.77 Class Average: 0.97 Class Bad: 1.00	0.7871
Random Forest	93.33	Class Good: 0.96 Class Average: 0.82 Class Bad: 1.00	Class Good: 0.84 Class Average: 0.96 Class Bad: 1.00	0.8337
SVM	84.7	Class Good:1.00 Class Average:0.48 Class Bad:0.00	Class Good: 0.52 Class Average: 0.96 Class Bad:1.00	0.5546

Table 4: Evaluation for classification with combined data

5.3 Experiment 3: Classification using K-fold Cross Validation

Since the dataset used for the classification of dishes is small in size, a widely used sampling method called K-fold Cross Validation is used, using which K-number of datasets are prepared using random cases from the original dataset and is used in the model of choice, whose results are then averaged later, to get a more accurate prediction value

(Duda et al.; 2000). The work by Rodriguez et al. (2010), tested cross validation experiments, with various values of k and is found that the value of k=10 gives accurate results. Hence, K-fold cross validation with K-value as 10 is used in order to create 10 random samples from the training data. Separate experiments are run where classification over data of a single restaurant and on combined data of 15 of the top restaurants.

- **Single Restaurant with 10 fold Cross validation**
  By using the K-fold cross validation training method with K-Value 10, the results are much better for the data from a single restaurant. From the below table, K-Nearest Neighbors with K-value 4 resulted in the highest accuracy, whereas surprisingly Random forest did not achieve high accuracy when using K-fold Cross Validation.

Model	Accuracy	Kappa
K-NN with k=4	97.5	0.9
K-NN with k=2	94.16	0.89
Random Forest	96.66	0.90
SVM	97.5	0.9

Table 5: Classification of single restaurant with 10 Fold cross validation

- **15 Restaurants with 10 fold Cross validation**
  A similar improvement can be seen by applying 10 fold Cross validation on the combined dataset of 15 restaurants. In this case, where the training dataset size was much larger than the previous experiment, where data from only one single restaurant was being used, Random Forest achieved in the highest accuracy, i.e. 94.99%. Again, Random Forest, did not achieve the best accuracy value when using K-fold Cross Validation method.

Model	Accuracy	Kappa
K-NN with k=6	90.4	0.75092
K-NN with k=9	90.3	0.74420
Random Forest	93.8	0.85007
SVM	94.99	0.87446

Table 6: Classification of combined data with 10 Fold cross validation

5.4 Conclusion

From the above experiments, it can be concluded that the proposed methods with various sentiment scores, Food Frequency, etc as the features, a classification model can be trained which results in a very highly accurate model, in order to distinguish the best dishes from the rest as proposed as the main aim of the research. The same is achieved across different experiments with various dataset sizes, which can be seen from the confusion matrices.

Since, the original dataset is of a very small size, the classification models are also trained and tested using 10 fold cross validation method as well, which gave even better results. Hence, it can be said that, the size of the data has not affected the final outcome of the experiments.

6 Discussion

This research aimed at recommending the best dishes served at a restaurant to a new customer, based on previous customer reviews. There have been many other studies which perform similar aspect based sentiment classification on customer reviews data, and achieved accuracy of 86% (Zhu et al.; 2016), 68% (Patra et al.; 2015). But by using a unique and effective approach, using modern text analytical tools and technologies, such as POS Tagging, Tokenization, Sentiment lexicon for sentiment score generation, a list of food dictionary used as seed words, and classification algorithms such as SVM, KNN and Random Forest, this research has successfully demonstrated that it is possible to achieve an accuracy of more than 97.5% to correctly classify food dishes into various classes.

Although the original reviews dataset is of more than 70K records combined for the top 15 restaurants from the *Yelp* ((Accessed July 28, 2018) dataset, the processed data that is passed to the classification models, is reduced to a little over 800 rows, which is a less than optimal size for training a multi class classification model. Even though by using 10 Fold Cross Validation technique, this issue of data scarcity is verified and is found to have no effect on the accuracy of the prediction model, the model can be further trained and tuned on a much larger dataset.

Restaurant reviews is just one of the domains, in which a customer can post reviews based on his/her experience. Many other domains such as e-commerce, Movie reviews or any other domain where product/services need to be recommended to a customer based on the previous customers' reviews can benefit from using such an implementation.

7 Conclusion and Future Work

This research presented a novel approach on distinguishing the food dishes served at restaurants by classifying the dishes into three classes i.e. Good, Average and Bad, based on the previous customer reviews.

Although this research classifies the food dishes highly accurately, some work still needs to be done, to get more relevant results. The system built during this research finds food dish names which are only one word long. So dishes such as "*Pizza*" are processed correctly, but dishes such as "*Margherita pizza*" are missed out. For this research, only reviews from *Yelp* ((Accessed July 28, 2018) are used. Future research could make use of combining reviews from multiple other review websites such as *Tripadvisor* ((Accessed July 28, 2018) or other popular sources, and comparing the results to find out which website has more useful and meaningful reviews.

This research is primarily focused on analysing restaurant reviews, but by changing the dictionary that is seeded for finding various dishes, to some other domain specific dictionary, a similar approach can be applied to future studies as well.

References

Akhtar, M., 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**: 116135.

Appel, O., Chiclana, F., Carter, J. and Fujita, H. (2016). A hybrid approach to the sentiment analysis problem at the sentence level, *Knowledge-Based Systems* **108**: 110124.

Ashok, M., Rajanna, S., Joshi, P. and Kamath, S. (2016). *A personalized recommender system using Machine Learning based Sentiment Analysis over social data*.

Bhattacharjee, K. and Petzold, L. (2017). What drives consumer choices? mining aspects and opinions on large scale review data using distributed representation of words, p. 908915.

Chinsha, T. and Joseph, S. (2015). A syntactic approach for aspect based opinion mining, p. 2431.

Dayan, A., Mokryn, O. and Kuflik, T. (2015). *A two-iteration clustering method to reveal unique and hidden characteristics of items based on text reviews*, p. 637642.

Ding, X., Liu, B. and Yu, P. (2008). *A holistic lexicon-based approach to opinion mining*, p. 231239.

Duda, R. O., Hart, P. E. and Stork, D. G. (2000). *Pattern Classification (2Nd Edition)*, Wiley-Interscience.

Esuli, A. and Sebastiani, F. (2006). *SENTIWORDNET: A publicly available lexical resource for opinion mining*, p. 417422.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). From data mining to knowledge discovery in databases, *AI Magazine* **17**(3): 37.

Garca-Pablos, A., Cuadros, M. and Rigau, G. (2018). W2vlda: Almost unsupervised system for aspect based sentiment analysis, *Expert Systems with Applications* **91**: 127137.

Guerini, M., Gatti, L. and Turchi, M. (2013). Sentiment analysis: How to derive prior polarities from sentiwordnet, p. 12591269.

Hossain, F. M. T., Hossain, M. I. and Nawshin, S. (2017). *Machine learning based class level prediction of restaurant reviews*, p. 420423.

Kang, H., Yoo, S. and Han, D. (2012). Senti-lexicon and improved nave bayes algorithms for sentiment analysis of restaurant reviews, *Expert Systems with Applications* **39**(5): 60006010.

Marcus, M. P., Marcinkiewicz, M. A. and Santorini, B. (1993). Building a large annotated corpus of english: The penn treebank, *Comput. Linguist.* **19**(2): 313–330.
URL: *http://dl.acm.org/citation.cfm?id=972470.972475*

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013). Efficient estimation of word representations in vector space, *arXiv:1301.3781 [cs]* .
URL: *http://arxiv.org/abs/1301.3781*

Mouthami, K., Devi, K. N. and Bhaskaran, V. M. (2013). *Sentiment analysis and classification based on textual reviews*, p. 271276.

Panchendrarajan, R., Ahamed, N., Sivakumar, P., Murugaiah, B., Ranathunga, S. and Pemasiri, A. (2017). Eatery - a multi-aspect restaurant rating system, p. 225234.

Patra, B., Mukherjee, N., Das, A., Mandal, S., Das, D. and Bandyopadhyay, S. (2015). Identifying aspects and analyzing their sentiments from reviews, p. 915.

Pealver-Martinez, I., Garcia-Sanchez, F., Valencia-Garcia, R., Rodrguez-Garca, M., Moreno, V., Fraga, A. and Snchez-Cervantes, J. (2014). Feature-based opinion mining through ontologies, *Expert Systems with Applications* **41**(13): 59956008.

Rodriguez, J. D., Perez, A. and Lozano, J. A. (2010). Sensitivity analysis of k-fold cross validation in prediction error estimation, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **32**(3): 569575.

Sauper, C. and Barzilay, R. (2013). Automatic aggregation by joint modeling of aspects and values, *Journal of Artificial Intelligence Research* **46**: 89127.

Sharma, A. and Dey, S. (2012). A document-level sentiment analysis approach using artificial neural network and sentiment lexicons, *SIGAPP Appl. Comput. Rev.* **12**(4): 6775.

Staiano, J. and Guerini, M. (2014). Depechemood: a lexicon for emotion analysis from crowd-annotated news, *arXiv:1405.1605 [cs]* . arXiv: 1405.1605.
URL: *http://arxiv.org/abs/1405.1605*

The Stanford Parser: A statistical parser ((Accessed July 28, 2018)). `https://nlp.stanford.edu/software/lex-parser.shtml`.

Trevisiol, M., Chiarandini, L. and Baeza-Yates, R. (2014). Buon appetito: Recommending personalized menus, p. 327329.

Tripadvisor ((Accessed July 28, 2018)). `https://tripadvisor.com/`.

Tripathy, A., Agrawal, A. and Rath, S. K. (2016). Classification of sentiment reviews using n-gram machine learning approach, *Expert Systems with Applications* **57**: 117126.

Wan, Y. and Gao, Q. (2016). An ensemble sentiment classification system of twitter data for airline services analysis, p. 13181325.

Wiebe, J. and Mihalcea, R. (2006). Word sense and subjectivity, *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, Association for Computational Linguistics, p. 10651072.
URL: *https://doi.org/10.3115/1220175.1220309*

WordNet ((Accessed July 28, 2018)). `https://wordnet.princeton.edu/`.

Yelp ((Accessed July 28, 2018)). `https://yelp.com/`.

Zhang, R., Zhang, Z., He, X. and Zhou, A. (2015). *Dish comment summarization based on bilateral topic analysis*, Vol. 2015May, p. 483494.

Zhu, Y., Moh, M. and Moh, T.-S. (2016). Multi-layer text classification with voting for consumer reviews, p. 19911999.