Market structure analysis for the restaurants serving Indian cuisine in Dublin using comparative opinions and clustering

MSc Research Project
Data Analytics

Ravi Kiran Halladakoppalu Shivanna
x17111609

School of Computing
National College of Ireland

Supervisor: Vikas Tomer

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

Student Name:	Ravi Kiran Halladakoppalu Shivanna
Student ID:	x17111609
Programme:	Data Analytics
Year:	2018
Module:	MSc Research Project
Lecturer:	Vikas Tomer
Submission Due Date:	17/09/2018
Project Title:	Market structure analysis for the restaurants serving Indian cuisine in Dublin using comparative opinions and clustering
Word Count:	5213

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:	17th September 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):	

Market structure analysis for the restaurants serving Indian cuisine in Dublin using comparative opinions and clustering

Ravi Kiran Halladakoppalu Shivanna
x17111609
MSc Research Project in Data Analytics

17th September 2018

Abstract

The restaurant industry is a very competitive one, everyone wants to be ahead of the others, but how? online reviews is the answer, they have abundant information from various people who visit these restaurants frequently comparing the services and giving out their opinions. So a method of identifying the comparative sentences using the restaurant names was used. A single line graph network was used to check for the direction and number of the comparisons for each restaurant. The market structure could be easily determined and visualized by this single line graph. The importance values were calculated based on the sentiment scores present in each restaurant, from which the focal restaurant could be easily identified using the cluster analysis. Also, this gives us the top restaurants in the market. Managers of the restaurants will get to know the direct and indirect competition from the other restaurants. They can keep an eye on the competitors and devise their course of action to be ahead of them.

1 Introduction

The players in every industry wants to be competitive and would like to be ahead of their competitors. Back in the days they used to get information about their competitors through surveys, newspapers and articles(Gao et al.; 2018). Nowadays, due to the advancement in technologies, like the invention of smart phones and smart devices, people are just a button click away from accessing the social media. Internet(Social media, Blogs, Forums) has become the mainstream communication channel these days. People discuss about the products or any of the services they have purchased or used. They tend to compare and contrast about the experiences they had and give their opinions about them.

Restaurant industry is not any different from the other industries. People who visit these establishments often write about their experiences in restaurant websites, their personal blogs or any third party websites such as trip-advisor, yelp etc in the form of reviews. However, these reviews will be in plain text and one cannot go through all of their opinions as it is time consuming. Also, if people want to compare and contrast between the restaurants then they have to keep on reading the reviews on a particular restaurant page which is out of scope. If a framework is developed that can identify the comparative reviews and portray which is a good restaurant would be helpful for the people to make better choices. Also, the manager of the restaurants can also have an eye on where their restaurant stands in the market, whether the users have a good opinion on them or not. Whom are they comparing us with, these types of questions can be answered.

We can at least find 10 percent of the total reviews that are comparative(Jindal and Liu; 2006a). So in this paper I am exploiting these type of reviews to find **”How the restaurants serving Indian cuisine in Dublin are placed in the market and who are their competitors, using comparative opinion mining”.**

Previous studies on the comparative opinions were conducted between two products, to find which product is better than the other. A multidimensional approach on comparing the restaurants based on the aspects were done recently(Gao et al.; 2018) but major works were done on the Chinese reviews and they have described that the Chinese language had more comparative words which is good for analysis. Most of the multi-dimensional analysis on the reviews were only on Chinese language, so in this paper extending the research to English language and instead of using the aspects, I am generalizing the analysis by using the keywords or restaurant names in the reviews to find that there is some comparison happening i.e. if there is a mention of the restaurant name other than itself, then there is some comparison in the review.

The data was scraped from trip-advisor and the restaurant names were used as keywords to identify the comparative reviews and the sentiment scores were calculated accordingly. Later they were represented as a network using graph analysis, the weight of the edges were calculated and were also used to calculate the importance of the nodes. Nodes were clustered according to their importance values. Section 2 explains about the previous works on comparative opinion mining, sentiment analysis and market structure analysis. Section 3 explains the methodologies used to carry out the analysis. Section 4 is about the implementation of the methods in R studio. Section 5 gives out the evaluation. Section 6 is all about the future work and conclusions of the research.

2 Related Work

This section is structured into three subsections. Identifying comparative opinions in subsection 2.1, Sentiment analysis in subsection 2.2 and Market structure analysis in subsection 2.3

2.1 Identifying Comparative Opinions

The first work on comparative sentences were done by(Jindal and Liu; 2006a) where they proposed the framework to identify comparative sentences. They have claimed that the earlier researchers were done only on the sentence constructs taking semantics and syntax into considerations. The authors also have proposed many methodologies in identifying the comparative sentences like POS tagging a sentences and extracting the sentences having tags superlative adjective/JJS, comparative/JJR/, comparative adverb/RBR and superlative adverb/RBS as they are the comparative adjectives and adverbs in a sentence. Using keywords to find the comparisons, they made a list of keywords that were manually identified in a comparative sentence and they were used to distinguish between the sentences. Other methods were the manual rules and the class sequential rules(CSR) with the multiple minimum support and using these with the combination of Naive Bayes and SVM machine learning techniques to classify as comparative and non comparative sentences.

After identifying the comparative sentences they have also discussed about how to mine these sentences from documents or blogs(Jindal and Liu; 2006b). Comparative sentences can be generally classified as *gradable* and *non gradable comparatives*. Both of them can be classified as follows, the first three are gradable and the last one is non gradable(Jindal and Liu; 2006b):-

1. Non-Equal gradable:- Whenever we find the comparison words such as "greater" or "lesser than" they are non equal gradable

2. Equative:- Where there is a comparison made equally on some features, using the word "equal"

3. Superlative:- These type of comparison contains the word "best", ranking it to the top of all the others

4. Non Gradable:- Where there is no clear distinction between the features in sentence or rather confusing sentence constructs *"Pen X and Pen Y are different in many ways*

Let us consider an ex:- "The food in Restaurant A is better than Restaurant B". Comparative relation is extracted as (relationWord, feature, entity 1, entity 2), so if we extract the relation from example sentence, it ca be written as (better, food, Restaurant A, Restaurant B). Also a 'type' was added later to distinguish between the gradable and non gradable comparatives.

Class sequential rules(CSR) and Label sequential rules(LSR) were used for identifying the gradable comparatives, Class sequential rules work like the pattern matching algorithms that was proposed by (Agrawal et al.; 1994). Label sequential rules was by using the wild cards to match the pattern with some predefines wild card like 'R*' that matches anything with this pattern will be returned. The *minsupport* and *minconfidence* is altered to get the best results. Overall, the LSR with keywords gave the best precession out of all the other methods used.

There was a shortcoming in the research by(Jindal and Liu; 2006b). They did not specify which entity to choose i.e. the entity that was preferred by the author. It was handled well in a research done by (Ganapathibhotla and Liu; 2008). They achieved the above problem by categorizing the comparative words found in the sentences as opinionated comparatives and context dependent comparative opinions. the comparative words were also categorized as increasing comparatives('more','longer' etc) and decreasing comparatives('lesser','fewer' etc), this was helpful in finding the entities preferred. Coming back to the opinionated comparatives, it's easy to find the entities if they had words like 'better','worse' etc. Words like 'more','most','less' and 'least' can be handled by the categorized groups of increasing and decreasing comparatives.

For the context dependent comparatives it is very much necessary to have the domain knowledge because there will be comparison like "The Scooter A gives more mileage per gallon than Scooter B". We need the domain knowledge of the words like 'mileage','gallon' etc to decode. These kind of the issues were effectively handled using the logarithmic functions to extract the preferred entities.

Xu et al stated that a sentence can contain multiple comparisons(Xu et al.; 2011) unlike the previous studies made by(Jindal and Liu; 2006a)(Ganapathibhotla and Liu; 2008). They used two level conditional random fields with a unfixed interdependencies for mining th comparative relations. A comparative relation was defined and denoted using the expression R(E1, E2, A, S). Where 'E1' is the first entity, 'E2' is the second entity, 'A' is the attribute, 'S' is the sentiment word and 'R' gives us the direction of comparison being made. The directions are >(better), <(worse), =(equal) and ∼ (no direction)

If we consider the example sentences, the relations can be written as follows:-

- *"Restaurant A has better food than Restaurant B"*
  > (Restaurant A, Restaurant B, food, better)

- *"Compared to Restaurant A, Restaurant B has good service"*
  < (Restaurant A, Restaurant B, service, good)

- *"Restaurant A and Restaurant B are both in good locations"*
  ∼ (Restaurant A, Restaurant B, location, good)

The sentences can have multiple comparisons and they can be written as follows:-

- *"Restaurant A has better food than Restaurant B and Restaurant C"*
  r1: > (Restaurant A, Restaurant B, food, better)
  r2: > (Restaurant A, Restaurant C, food, better)

- *"Restaurant A has good environment but higher price than Restaurant B"*
  r1: > (Restaurant A, Restaurant B, environment, good)
  r2: < (Restaurant A, Restaurant B, price, high)

Graphical methods were used to evaluate the results. Two way graphs were used with unfixed interdependencies. Also, they have compared SVM, custom random fields with, without dependencies and evaluated results.

Another method was proposed by (Tkachenko and Lauw; 2014) where the comparative relations were mined at the sentence and entity levels. Instead of finding the comparative words they used products names in the dictionary list to find the comparative relations, this was called the Named Entity Recognition(NER). A *CompareGem* model was also introduced, the algorithm takes Gibbs sampling into consideration. It was the best choice to handle the uncertainties that is found in the text by using the probabilistic approach. As it would consider entity and sentence it was call a 'joint model'. And it is generative as it could handle both supervised and unsupervised learning. Also, the *CompareGem* model gives the best results for the sentence level comparative sentences(Tkachenko and Lauw; 2014)

All the earlier studies like (Jindal and Liu; 2006a)(Ganapathibhotla and Liu; 2008) were concentrated on the comparative predicates, the words near to the products that express comparative relation. CSR's were used to mine the comparative relation, they need the comparative predicates and they had windowing effect, very limited to catch the long range dependencies. So a new method was introduced call the skip node kernel method. It is a tree structure method and works on exact computations or any approximate computations. There were many other tree structure methods like the sub tree, partial tree and sub set tree. But they had shortcomings and ambiguity in calculating the similarity values so skip node was proposed(Tkachenko and Lauw; 2015)

A review was conducted on all the research and methodologies that were carried out on comparative mining by (Varathan et al.; 2017). The clear distinction between the techniques were made as, 'Machine learning', 'Rule mining' and 'Natural Language Processing'. There is a statement made in this paper saying that all the previous researches on comparative mining were on products . Nothing was done on restaurants, bus services and political figures(Varathan et al.; 2017)

A recent study was done on comparative relation mining on Chinese reviews to check the competitiveness between two restaurants. The Chinese sentence constructs are different and contain many comparative adjectives(Wang et al.; 2017). So between two restaurants 7 aspects such as 'quantity of dishes','ease of reservation','service','environment','price' and 'taste' were taken into account and the restaurants were compared based on these factors. This research was extended to 50 restaurants that serve Schezwan cuisine to find the competitors among them using the single line, dichotomic and multi line graph analysis to check for the focal restaurant which is sets standard for all the other restaurants in the market(Gao et al.; 2018).

2.2 Sentiment Analysis

Opinion mining or sentiment analysis is nothing but users feeling mentioned in the form of text as positive, negative or neutral. It was done on the extracted features of a particular products ex:- Digital camera, the 'lens' and 'size' are some of its features. The positive, negative and neutral scores were extracted from the reviews to get the user sentiment towards the product(Hu and Liu; 2004). This was developed into a package 'sentimentr' in 'R' programming to calculate the sentiment scores of the sentences in reviews, the scores ranged between -1 to +1. Until recently an extension of this was proposed by(Jockers; 2018) in that he has mentioned if any adversative conjunctions('although','even though' etc) appear near the polarized words, then the scores can we altered and can take the values more than the specified range of -1 or +1.There is a detailed explanation on this in a github page [1]. Also, the sentiment analysis can be done on many levels like document, sentence and word level(Wang and Wang; 2014). It depends on the research to choose which type of analysis is suitable to find the perfect solution.In their related word they have explained about the authors that have worked on document, sentence and word levels.

Another research was done lately where the reviews from tourism industry were considered and they were given the polarity scores and classified using the Naive Bayes and decision trees(Songpan; 2017). They have evaluated and visualized the positive and negative words used in constructing the sentences.

2.3 Market Structure Analysis

Previous studies on the market structure and competitor analysis was done on the resource similarity and market commonality(Chen; 1996).In which he had defined *"The degree of presence that a competitor manifests in the markets it overlaps with the focal firm"* and *"The extent to which a given competitor possess strategic endowments comparable, in terms of both type and amount to those of the focal firms"*

Resource and market structure based frameworks were used to identify the competitors, manager of the firms will always focus on the focal firms but they don't know that

[1] https://github.com/mjockers/syuzhet

there is a possibility of getting competition from unexpected quarters(Peteraf and Bergen; 2003). Many complex ways of identifying the competitors and the market structure were discussed in (Wu and Olk; 2014)(Peng and Liang; 2016). they have used several survey models to get the results on the market structure. There are no other works on the market structure based on the comparative opinions.

3 Methodology

The methodology that is proposed in this paper is the knowledge discovery in database(KDD) because the data used for the research is unstructured data in the form of reviews(Fayyad et al.; 1996). Machine learning techniques are very useful for the analysis, there are two types of the machine learning techniques called the supervised and unsupervised learning. In my research unsupervised technique is used for clustering the nodes according to their importance values. Figure 1 will be a representation of the KDD process and Figure 2 is the implementation of KDD process.

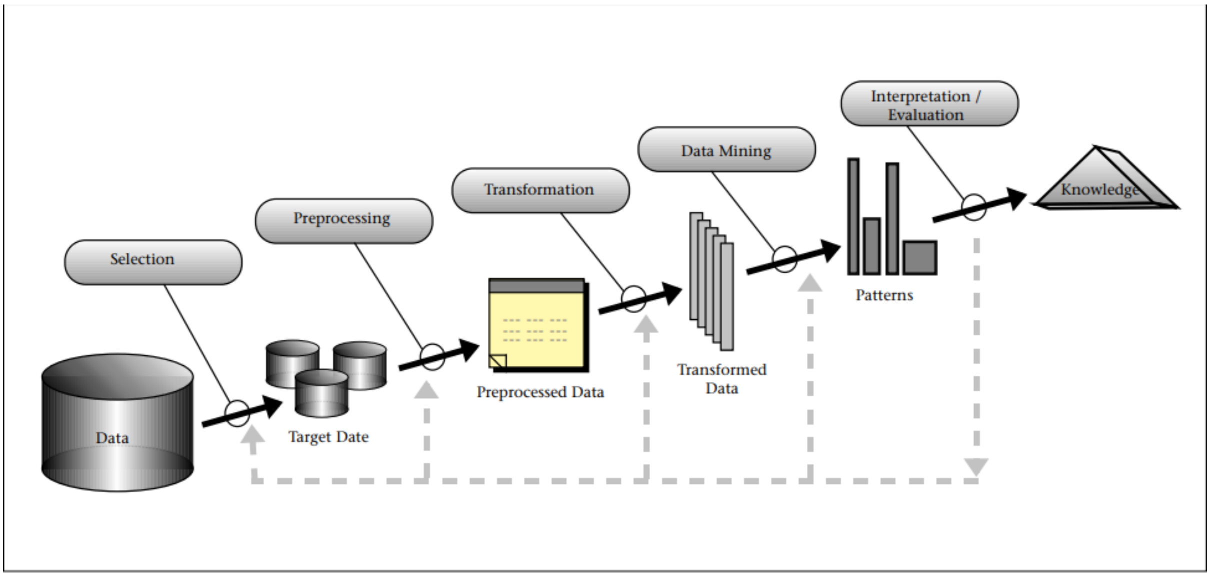

Figure 1: KDD process (Fayyad et al.; 1996)

3.1 Data selection, crawling and storing

Shortlisting the data for analysis was a difficult task as the research question was finding the market structure of the restaurants that serve Indian cuisine in Dublin. After going through a lot of websites the data from *TripAdvisor* was chosen.Using the filters, restaurants serving Indian cuisine in Dublin were considered. A web crawler was developed using the 'rvest' package in R programming to extract the reviews and stored them as data frames by the restaurant names. A total of 80 restaurants were found, but only reviews from 76 restaurants were extracted. the other 4 did not have any reviews written. The restaurants are renamed as P1, P2....P76.

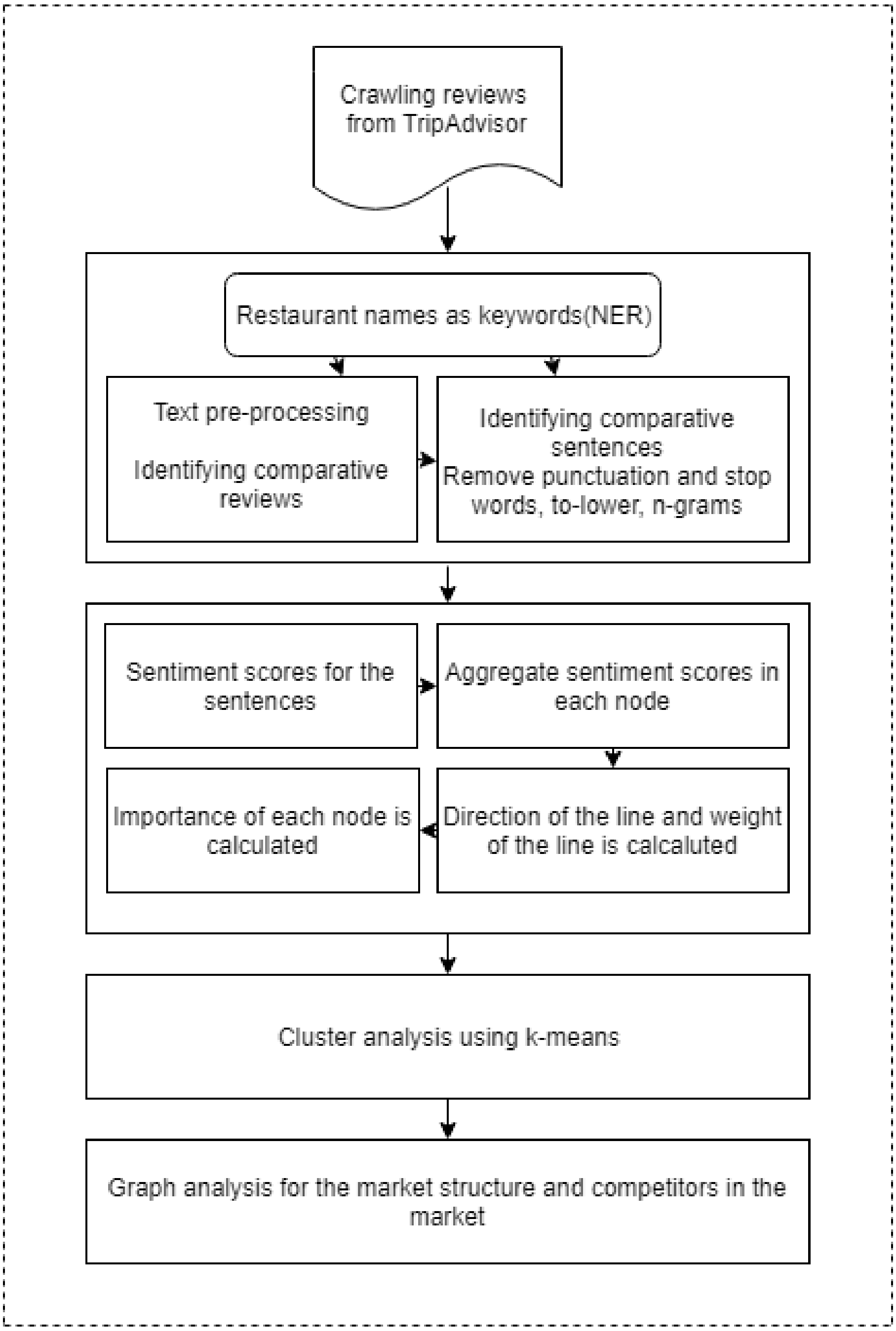
Crawling reviews from TripAdvisor
Restaurant names as keywords(NER)
Text pre-processing
Identifying comparative reviews
Identifying comparative sentences
Remove punctuation and stop words, to-lower, n-grams
Sentiment scores for the sentences
Aggregate sentiment scores in each node
Importance of each node is calculated
Direction of the line and weight of the line is calcaluted
Cluster analysis using k-means
Graph analysis for the market structure and competitors in the market

Figure 2: Implementation of KDD process

3.2 Data Pre-processing

For the analysis only review columns were required, only they were selected and stored as vectors. These were later put into a list. Not much cleaning was required as all the reviews were in order and high quality. The restaurant names were used as the keywords, if a restaurant name is mentioned in any of the reviews there is a high possibility of comparison. This method was used instead of the POStagging and extracting reviews that have comparative adjectives and comparative adverbs. It was used to handle the words such as 'although','in spite of' etc are some of the example words used in sentence structure for comparing which will be missed in POStagging. So, in this research if a restaurant name is mentioned in a review it is perceived as a comparative sentence.

The reviews of each restaurant that was stored in a list was converted into a corpus of documents. After that the cleaning of text was done by removing punctuation, converting to lowercase, n-grams were used to split the sentences into keywords (Damashek; 1995) and the document term matrix was used to filter out the reviews that have the restaurant names(Mouthami et al.; 2013). Once again the same process was repeated at a sentence level, only the sentences having restaurant names were retained. At the end stop words were removed and reviews were ready for transformation.

3.3 Data Transformation

The sentiment scores of the sentences are computed, scores can be positive or negative. Now we have the data set having three columns, restaurant names in the first column, the restaurant that are mentioned in the reviews and sentiment score of the sentence(Hu and Liu; 2004). As mentioned in the related work, scores can be from -1 to +1 and these are altered in the presence of conjunctions(Jockers; 2018).

After this to build the network sentiment polarity scores are aggregated between the two nodes to find out which is the strongest node out of the two nodes. Lets consider two nodes P1 and P2, equation (1) is used to aggregate the scores between nodes P1 and P2(Gao et al.; 2018).

$$\sum_{i=1}^{n(i)} Score_i \tag{1}$$

Where
n(i) = number of pairs
$Score_i$ = comparative result of i^{th} pair

If the score is > than 0 then the line is drawn from P1 ⟶P2. if it is < 0 then the line is drawn from P2 ⟶ P1. if it is '0' then they are just connected by a line with no direction(Wang et al.; 2017). Then weight of the line is determined by equation (2)(Gao et al.; 2018)(Wang et al.; 2017).

$$WeightP_1P_2 = \frac{1}{n(i)}[\sum_{i=1}^{n(i)} \{Score_i\}_{P_1P_2}] \tag{2}$$

Where
n(i) = number of pairs
$\{Score_i\}_{P_1P_2}$ = i^{th} comparative result between P1 and P2

Finally the importance values of each node is calculated using the equation (3)(Gao et al.; 2018).

$$C_D(n_i) = [\sum_j x_{n_i n_j}]/(N-1) * [N^+_{n_i}/N^-_{n_i}] \tag{3}$$

Where
N = number of the nodes
$x_{n_i n_j}(x_{n_j n_i})$ = compassion between two nodes n_i, n_j
$x_{n_i n_j} = x_{n_j n_i} = 0$, There is no direct relation between the two nodes
$x_{n_i n_j} = x_{n_j n_i} = 1$, There is definitely a relation between the two nodes
$[\sum_j x_{n_i n_j}]$ = Total amount of all the direct relations
$N^+_{n_i}$ = +ve comparison pairs in node n_i
$N^-_{n_i}$ = -ve comparison pairs in node n_i

3.4 Data Mining

After all the transformation we will be left with the nodes and their importance values. Using the unsupervised machine learning we can check how the nodes are placed in the clusters. For the purpose of clustering I am using the K-means clustering technique that was proposed by(Hartigan and Wong; 1979). To check the ideal number of the clusters many values will be given to 'k' and plotting these k values in a graph we can know the ideal value of 'k'using the elbow method (Thorndike; 1953). Accordingly those number of clusters are considered for further evaluation.

3.5 Visualization, Interpretation and Evaluation

For the visualization I need to draw a network between the nodes based on the aggregated scores discussed in the data transformation section ans this can be achieved by the methods mentioned in(Wang and Wang; 2014) (Wang et al.; 2017)(Gao et al.; 2018). Nodes are connected in a network to visualize the relations that are leaving and entering them. After the importance value are calculated using equation(3), the nodes with higher importance values are considered for a high level analysis. The node with highest importance value is the focal restaurant(Gao et al.; 2018).

A concept of geo desic distance is explained in (Wang and Wang; 2014), it can be defined as *"Geo desic distance is the shortest path from other nodes to the target node according to the link between two nodes"*(Gao et al.; 2018). Once the focal restaurant is found, there is a direct competition between restaurants if geo desic distance is '1', if it is '2' then there is no direct competition but there is some competition. If its '3' and more than it, then it's difficult to say there is competition(Varathan et al.; 2017)

4 Implementation

In the methodology section all the methods that were used from extraction to analysis are explained. In this section there will be a brief explanation on how it was done.

4.1 Crawling for reviews

A web crawler was developed using the 'rvest' package in R programming to extract all the reviews from restaurants serving Indian cuisine in Dublin from TripAdvisor. The

URL of each restaurant was given as input to the crawler to extract the reviews. Like this, reviews from the 76 restaurants were extracted and stored as a data frame. The data set consists of 4 columns 'reviewer id', 'quote', 'review' and 'rating'. Rows vary from restaurant to restaurant, while if take the data set as a whole it has 7824 rows. Few of the restaurants had two branches like 'bombay pantry fairview' and 'bombay pantry rathgar', they were merged as one for this analysis. One of the restaurant name was 'Pickle', so when used as a keyword it used to return the sentences containing dish and restaurant names. To deal with this problem, the sentences that were containing 'Pickle' as dish name were identified and deleted manually.

4.2 Identifying the comparative reviews

Only the column that had reviews was extracted from each hotels data frame and named as P1,P2 ... P76. Later they were stored into a list. Keywords for each restaurant is identified and stored into a dictionary. Restaurants name with single word are much easier, as we take their names directly. The restaurant names having two or more number of words is difficult to identify as people tent to use only the first word in their names ex: "Meghna Tandoori" is mentioned as only Meghna. To tackle this a list of names as keywords were made and they as also called Named Entity Recognition(NER)(Tkachenko and Lauw; 2014), when dealing with the product names, here our product names are our restaurant names. So these are identified and stored as a dictionaries according to their number of words. All the single words in one, two in another and so on until all the combination are done.

4.3 Text mining

All the reviews are now converted into corpus and using the text mining(tm) package in R programming. Then the punctuations are removed and the text is converted to lower case for the uniformity. Using 1-gram for the first time the words are tokenized, document term matrix is used with the dictionary that contain the single words as the keywords to filter out the reviews those have the names of restaurants. There is a high possibility that the user while giving the opinions they mention the same restaurants name in the reviews. ex: "Ananda ha good food" under the restaurant 'Ananda' itself, to exclude these kind of the reviews a function was developed to remove the common terms and the frequency is set. The above steps were repeated again with 2-grams, 3-grams, 4-grams with their respective dictionary list. Only the reviews that contain the restaurants were filtered out and kept.

The reviews were too long so it was split into sentences and using the same process of document term matrix with the dictionary list only the sentences with the restaurants names mentioned were retained.

4.4 Sentiment Polarity

The sentiment polarity score of each sentence is calculated by removing the stop words and using the' sentimentr' package in R programming. Finally the data set is represented as a tuple of 3 P1,P6, 0.25. P1 is the restaurant, P2 is the restaurant mentioned in the review and 0.25 is the sentiment score of the sentence. In table 1 an example is provided.

Restaurant	Restaurant Compared	Sentiment
P1	P6	0.250217297
P5	P6	0.000000000
P7	P9	0.000000000
P7	P9	0.1296362
P7	P62	0.5185450
P7	P9	0.4776679

Table 1: First 6 rows are shown from the dataset having sentiment scores

If the score is >0 according to the equation (1) a line is drawn from P1 to P2 , if it is <0, then the line will be from P2 to P1. if it's $=0$ then, they are just connected by a line with no direction

The sentiment scores are aggregated according to the equation (2) to get weight of the edges i.e. all the scores that exists between two nodes are taken and the average score is computed. In table 2 an example is provided.

Restaurant	Restaurant Compared	Avg Sentiment
P1	P6	0.250217297
P5	P6	0.000000000
P7	P20	0.125819890
P7	P27	0.558214286
P7	P33	0.053300179
P7	P54	0.023333333

Table 2: First 5 rows are shown from the dataset of grouped sentiment scores

Importance values are calculated using the equation (3). The nodes having both the positive and negative values are selected as per the equation and computed. In table 3 the restaurant and their importance vales can be seen.

Restaurant	Importance
P9	8.567373998
P14	0.446277340
P16	1.481457624
P37	0.169418437
P55	0.761590076
P60	0.053097407
P68	0.099062691
P70	-0.009502622
P73	1.383832664
P74	0.125847571
P6	2.287161138

Table 3: Restaurants and their importance values

4.5 Clustering

After the computation of importance values, they are clustered using the unsupervised machine learning algorithm 'K-means'. To find out how the restaurants are dividend among the clusters, 'elbow' method is used to find out the optimum number of the clusters.

5 Evaluation

The evidence for steps in implementation and hypothesis is provided in this section. Graph analysis is done to show how the network is connected, clustering show how the restaurants are placed in the clusters. Restaurant with the highest value is taken as the focal restaurant and the nodes directly connected is again visualized in the graph.

5.1 Market structure using single line graph

Only for 45 out of 76 restaurants serving Indian cuisine had comparative opinions available. The nodes are connected in a network using the directions given by the equation(1) and the values are from the table (1), they can be seen in figure 3. The lines with arrows pointing towards other nodes means that it is preferred over the other restaurants. Lines with the arrows pointing towards a node means that the other restaurant is preferred over the current one.

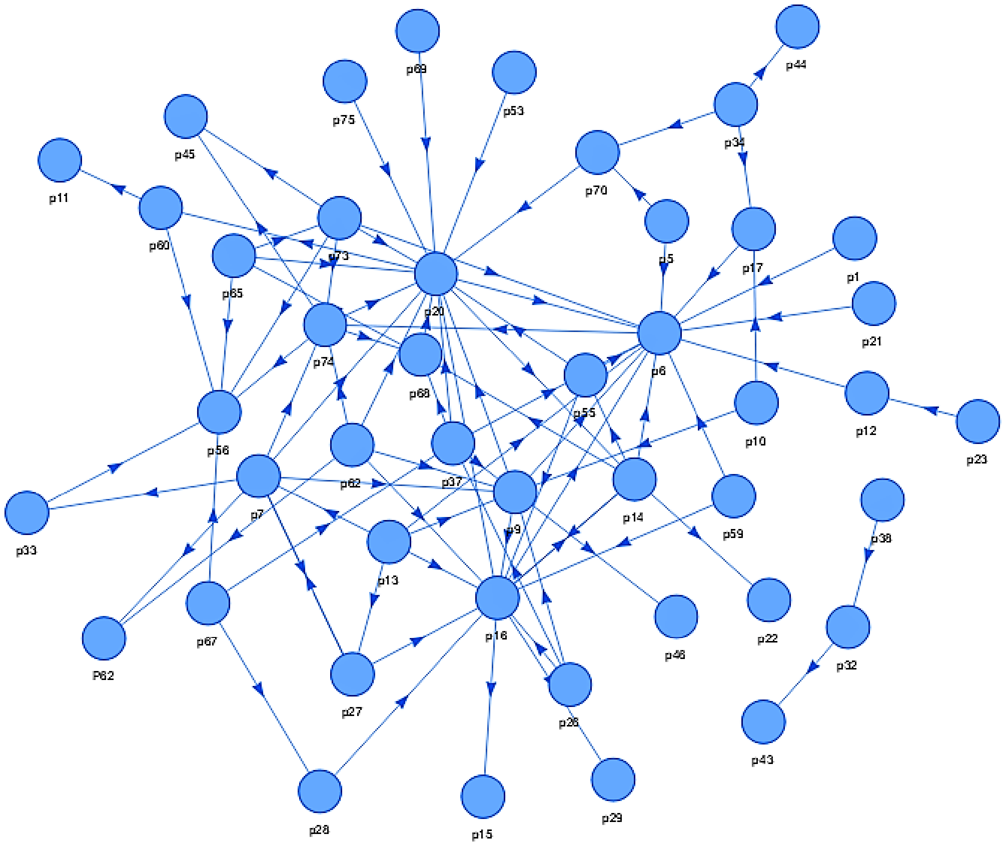

Figure 3: Single line graph of 45 nodes

5.2 Cluster Analysis using K-means

Out of the 45 nodes that had comparative opinions, only 11 of them have their importance values according to equation(3). The values in table 3 are clustered. To get the ideal number of the clusters elbow method was used and the ideal value of 'k' from the graph is '3', it can be seen in figure 4.

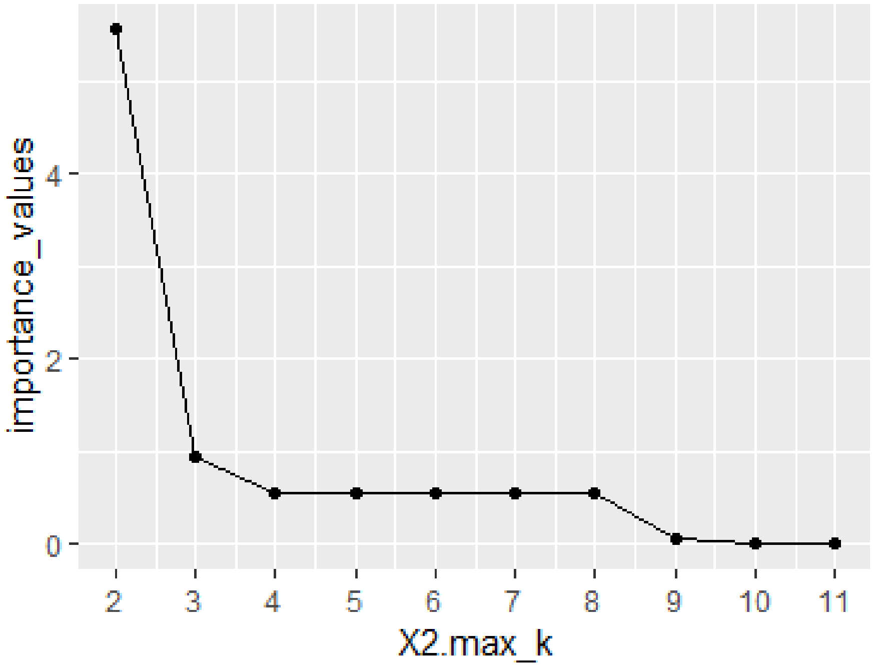

Figure 4: Elbow method graph

So the 'k' value is computed as '3', which results in 3 clusters. The visualization of clusters can be seen in figure 5.

So the first cluster has only one restaurant 'P9', with the highest importance value therefore it is regarded as the focal restaurant. Second cluster has seven restaurants, with 'P55' being the representative. Third cluster has 3 restaurants with 'P6' being the representative.

5.3 Competitors respective to nodes

Any node can be considered to check for it's competitors, I am using the focal node 'P9' to check for its competitors. Nodes that are connected directly to the 'P9' are direct competitors i.e the geo desic distance is '1', the nodes that are placed in the geo desic distance of '2' are the indirect competitors or it can be described as the there is some competition. Figure 6 shows the representation of 'P9' competitors

5.4 Discussion

From the reviews, only that had comparison were considered. Unstructured data was converted to have three columns, restaurant, the restaurant that is mentioned in the reviews and sentiment score of the sentence where restaurant name is mentioned. The

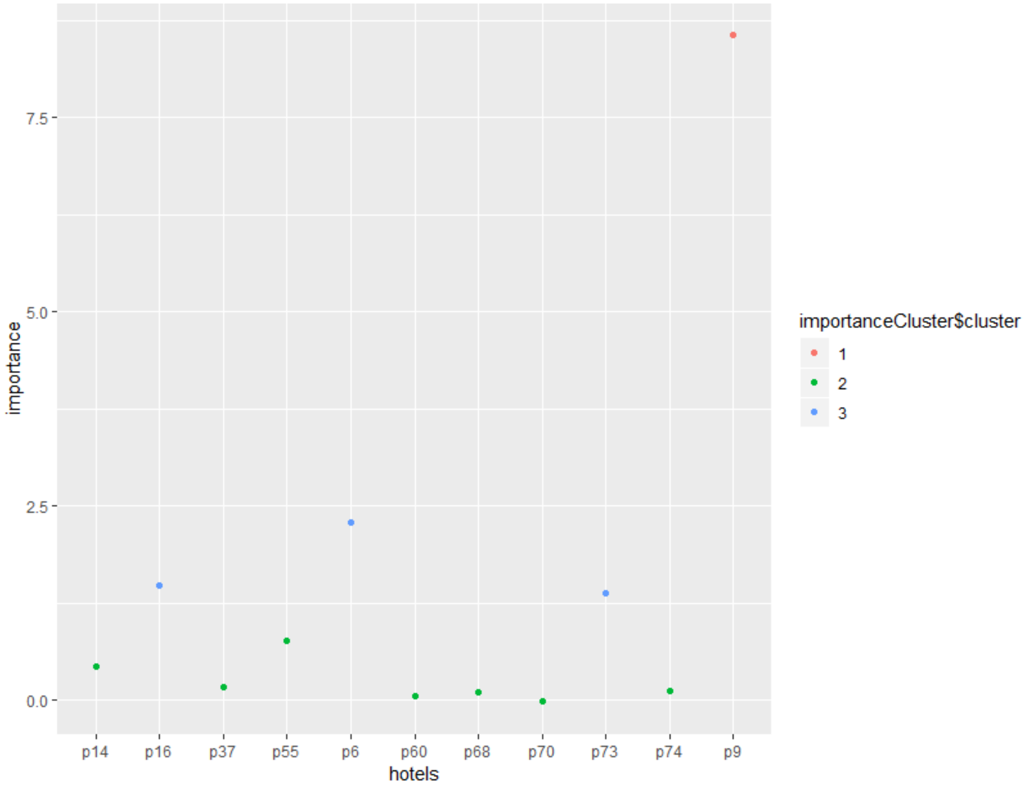

Figure 5: Cluster Analysis

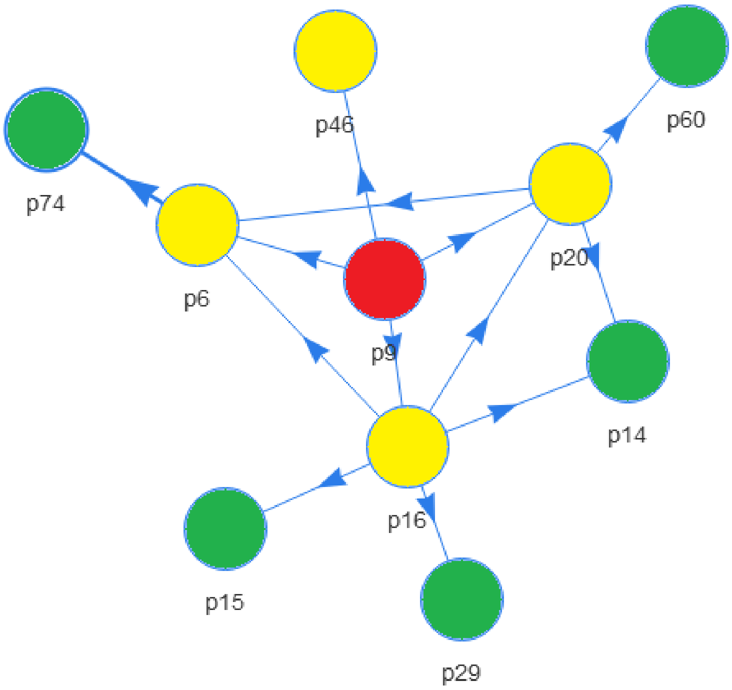

Figure 6: 'P9' and its competitors in the market

number of lines that are leaving the nodes represent that it is preferred over the others that it is entering into. Lines coming from the other nodes state that they are preferred over the one entering. Like this we can analyze the entire network of restaurants that are serving the Indian cuisine in Dublin.

The importance values as the name, it gives us how often it was used to compare with other restaurants. To calculate this value the node or restaurant should have both positive and negative sentiment scores. Higher the importance values, then they were often used for comparison. 'P9' has the highest value of '8.56' and that is a single restaurant in a cluster 1, it has outnumbered all others in the market. Next best importance value is of the node 'P6' with '2.28', the cluster 3 which it came from had three nodes and it is representative of the cluster 3. Cluster 2 has 7 nodes and 'P55' is its representative with value of '0.76'.

The importance value of the 'P9' is higher than the others because of the strong sentiment scores of the sentences. 'P9' is the focal restaurant in the market and it's immediate competitors are the restaurants or nodes marked yellow in the figure 6, they are 'P6','P16',P20' and P46'. There is an indirect competition from the nodes that are marked in green, they are 'P14','P15','P29', 'P60' and 'P74'.

The data set was small around 8000 rows and all the available reviews were extracted, some had 20 and some 500. As the restaurant started their operation in different time period, considering the time range between the reviews extraction will give better results.

6 Conclusion and Future Work

The research in this paper was all about the market structure analysis and the competitor of the restaurants using comparative opinions in the reviews. Using networks built, now the manager will know which of the restaurants are giving direct and indirect competition to his restaurant. He can follow up on these restaurant's services and improve themselves.

Customers will get to know the top rated restaurant, which one's to visit first and so on make the subsequent visits. Only the keywords or restaurants names were used to identify the comparative opinion i.e. if a restaurants name is present in the review, it means they are comparing something in it. Improving on these keywords should be made for better results, also a second round of filtering should be done using the POS tagging looking for the tags containing comparative adjectives so we can find the comparative reviews using classification.

In this research the restaurants having branches were considered as single restaurant by combining their reviews, further work needs to be done on distinguishing them. Quotes and ratings of the comparative sentences can also be used for the analysis, Quotes can be used as for text mining or the ratings for classification.

This can also be extended to any cuisine in the market, popular one to try on is the Irish cafe's serving breakfast in Dublin as there are many of them available in TripAdvisor with abundant reviews. Not only restaurants this text mining methodology can be extended to numbers other domains where multiple products are compared.The coding done on this needed manual runs by changing the values to pre-process the data, automation of this process can be taken under future work

7 Acknowledgment

I would like to thank my supervisor Mr. Vikas Tomer, examiners Mr. Vikas Sahani and Dr. Paul Stynes for their valuable inputs on the project. To my beloved family, without their support I couldn't have reached this far.

References

Agrawal, R., Srikant, R. and others (1994). Fast algorithms for mining association rules, *Proc. 20th int. conf. very large data bases, VLDB*, Vol. 1215, pp. 487–499.

Chen, M.-J. (1996). Competitor analysis and interfirm rivalry: Toward a theoretical integration, *Academy of management review* **21**(1): 100–134.

Damashek, M. (1995). Gauging similarity with n-grams: Language-independent categorization of text, *Science* **267**(5199): 843–848.

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

Ganapathibhotla, M. and Liu, B. (2008). Mining opinions in comparative sentences, *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, Association for Computational Linguistics, pp. 241–248.

Gao, S., Tang, O., Wang, H. and Yin, P. (2018). Identifying competitors through comparative relation mining of online reviews in the restaurant industry, *International Journal of Hospitality Management* **71**: 19–32.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm, *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **28**(1): 100–108.

Hu, M. and Liu, B. (2004). Mining opinion features in customer reviews, *AAAI*, Vol. 4, pp. 755–760.

Jindal, N. and Liu, B. (2006a). Identifying comparative sentences in text documents, *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 244–251.

Jindal, N. and Liu, B. (2006b). Mining comparative sentences and relations, *AAAI*, Vol. 22, p. 9.

Jockers, M. L. (2018). *syuzhet: An R package for the extraction of sentiment and sentiment-based plot arcs from text.*
URL: *https://github.com/mjockers/syuzhet*

Mouthami, K., Devi, K. N. and Bhaskaran, V. M. (2013). Sentiment analysis and classification based on textual reviews, *Information communication and embedded systems (ICICES), 2013 international conference on*, IEEE, pp. 271–276.

Peng, Y.-S. and Liang, I.-C. (2016). A dynamic framework for competitor identification: A neglecting role of dominant design, *Journal of Business Research* **69**(5): 1898–1903.

Peteraf, M. A. and Bergen, M. E. (2003). Scanning dynamic competitive landscapes: a market-based and resource-based framework, *Strategic management journal* **24**(10): 1027–1041.

Songpan, W. (2017). The analysis and prediction of customer review rating using opinion mining, *Software Engineering Research, Management and Applications (SERA), 2017 IEEE 15th International Conference on*, IEEE, pp. 71–77.

Thorndike, R. L. (1953). Who belongs in the family?, *Psychometrika* **18**(4): 267–276.

Tkachenko, M. and Lauw, H. (2015). A convolution kernel approach to identifying comparisons in text, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Vol. 1, pp. 376–386.

Tkachenko, M. and Lauw, H. W. (2014). Generative modeling of entity comparisons in text, *Proceedings of the 23rd acm international conference on conference on information and knowledge management*, ACM, pp. 859–868.

Varathan, K. D., Giachanou, A. and Crestani, F. (2017). Comparative opinion mining: a review, *Journal of the Association for Information Science and Technology* **68**(4): 811–829.

Wang, H., Gao, S., Yin, P. and Liu, J. N.-K. (2017). Competitiveness analysis through comparative relation mining: evidence from restaurants online reviews, *Industrial Management & Data Systems* **117**(4): 672–687.

Wang, H. and Wang, W. (2014). Product weakness finder: an opinion-aware system through sentiment analysis, *Industrial Management & Data Systems* **114**(8): 1301–1320.

Wu, J. and Olk, P. (2014). Technological advantage, alliances with customers, local knowledge and competitor identification, *Journal of Business Research* **67**(10): 2106–2114.

Xu, K., Liao, S. S., Li, J. and Song, Y. (2011). Mining comparative opinions from customer reviews for Competitive Intelligence, *Decision support systems* **50**(4): 743–754.