THE IMPACT OF ONLINE REVIEWS ON CUSTOMER BEHAVIOUR AND USAGE PATTERNS

A.N.K. Angulgamuwa

179304X

Degree of Master of Science in Computer Science

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

May 2020

THE IMPACT OF ONLINE REVIEWS ON CUSTOMER BEHAVIOUR AND USAGE PATTERNS

A.N.K. Angulgamuwa

179304X

Dissertation submitted in partial fulfillment of the requirements for the degree Master of Science in Computer Science specializing in Data Science, Engineering and Analytics

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

May 2020

DECLARATION

I declare that this is my own work and this thesis does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books)

Signature: ……..... Date:…….

Name: A.N.K Angulgamuwa

The above candidate has carried out research for the Masters under my supervision.

Signature of the supervisor: …… Date:……..

Name: Dr. Charith Chithraranjan

ACKNOWLEDGEMENTS

First, I’m grateful to Dr. Charith Chitraranjan for giving me the opportunity and further guidance in selecting and conducting this research. His continuous supervision greatly helped me in keeping the correct phase in research work. I especially appreciate the frequent feedback on the report, which helped me to correct and fine-tune it to this level. Last but not least, my heartfelt gratitude goes to my family and friends who supported me throughout this effort

Abstract

Online review forums and websites are highly popular these days. They enable customers to post online reviews and rate businesses based on their personal experience. These online reviews affect future customer decisions and demands on business. The Influence of the review might be high or low according to its user profile, overall image of the business and the context of the review itself. A well reputed or related user can add more weight to future customer decisions. A business with a popular brand name might not get rejected due to some negative comments. Also these reviews might start a trend on customers visiting that business or leaving that business. The main objective of this research is to predict customer behaviour for a given time period after a certain date using features of previous online reviews. Further to identify trends in customer behaviour and derive trending review topics that persist those trends.

Keywords: online reviews, multi class classification, change point detection, trending topics, frequent itemset mining

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATION

AUC	area under the curve
LR	logistic regression
GBDT	gradient boosted decision tree
SVM	support vector machine
CBC	Choice-based conjoint
MSE	mean squared error
RMSE	Root mean squared error
GM	Geometric mean
RUS	Random undersampling

LIST OF APPENDICES

1. INTRODUCTION

Consumers are always looking for feedback before purchasing a product no matter which era they live in. Nowadays everything is on a virtual platform that we can find well enough reviews on the internet. There are websites which are dedicated to maintaining customer reviews on businesses such as Tripadvisor, Yelp and Zomato. These online reviews can play a major role in customer behaviour in a positive manner or a negative manner. Or else review can have a neutral effect on customer decisions. If reviewers have a bad personal experience with a business they will write some negative comments. On the other hand when reviewers have experienced a pleasurable service they will write positive comments in favour of the business. So the content of the reviews can be positive or negative towards a business or service.

But only the idea behind a review can not affect future business alone. The one who posted that review is a concern. If the person has a reputed profile others may pursue his opinion than a review from a random profile. Same goes with a review that is written by an acquaintance. A profile with a good recognition can be identified. They may be experts in the area. They get more responses for their posts. They post so often in networks and have a long active history. The fan base of the writer can be large and the person can have strong connections to other reputed profiles within the network.

Most of these sites show the average ratings given by reviewers. And the overall rating of a reviewer depends on their requirement. When it comes to restaurants, if the parking area is a major concern for a reviewer, no matter how quality food that the restaurant serves the reviewer gives bad ratings and comments. It affects the average rating shown in the site. Restaurants can be categorized according to the services they give. Some of those categories are ethnic, fast food, fine dining and cafe. For each category there are unique factors. For ethnic restaurants they should be specialized for certain cuisine and their food should be able to compete with non specialized shops. For fast food restaurants food quality may be more important than the dress code. But fine dining restaurants expect guests to follow a dress code. For cafes ambience and environment is a more concern as the customers expect to spend quality time alone or with friends. If the requirements expected from the restaurant based on it’s category is fulfilled, reviewers will give them good ratings regardless of other lacking facilities.

Business reputation is also another factor when choosing a business. A company with a popular title already has a customer base which has grown around their brand name regardless of the online reviews. There might not be an effect on customer behaviour for that business. The usage patterns of consumers can be based on their advertising skills and marketing campaigns.

When studying the impact of user online reviews on user behaviour, it is better to take user online activities into the account rather than considering the external factors like revenue. So we can count on online user check-ins and reviews written after a certain review. The services of the business can evolve over time. Latest records should be focused when deriving insights from data as predictions should be based on the current state of the business.

The main objective of this research is to predict future customer behaviour using the features of business, reviewer and review text. Features of the reviewer will be explored thoroughly. A classification approach will be implemented to predict the reviews are increasing or decreasing the customer visits. Further, The trends of customer behaviour will be derived using time series segmentation techniques and trending topics that cause to persist those trends will be analyzed using frequent itemset mining process.

1.1 Yelp Data Set

1.1.1 Yelp.com

‘Yelp.com’; founded in 2004 is a fast growing website where voluntary consumers can leave their comments on services given by businesses. It has a large number of restaurant details as well as other businesses. Records says it has over 10 million business reviews and approximately 40 million unique visitors per month [2]. Registered users have free accounts in yelp that they use to write reviews. Yelp enables users to rate any business from one to five. These reviews and ratings are publicly available on the website for reading. Readers will learn about each business through context of the reviews.

But it is not the first place users can view the reviews. Most of the time they go through

a search process which includes some filtration according to the users requirements. Ratings, category and location are some filters offered by yelp. After that readers can look for other features like parking, allow take away, accept credit cards, dress code etc. Even though the business has competitive ratings and reviews, if it lacks some features that a particular customer is looking for, the customer will not consider that business.

1.1.2 Yelp Challenge

Yelp launches data science challenges every year and publishes their datasets on the internet. These data sets are publicly available and allowed to be used for research purposes. There are a lot of research papers which use yelp data sets and most of them are researching user reviews. Also there are many papers written about consumers' perception about online reviews using other data sets. The research conducted by using yelp data will be further discussed in the literature review.

1.1.3 Yelp Data

Yelp Data set consists of details about businesses, reviews, users, online check-ins, tips and photos. Business details include information such as facilities provided by the business, geographical information, business category, opening hours and current ratings given by yelp users. Yelp user's history, posts, received comments, fans and friends are included in user details. Both tips and reviews express the users experiences about a business. Tips are short comments which contain some key information about a business. Review is a more descriptive comment where the user writes the whole experience. Reviews can contain multiple paragraphs where tips contain one or two sentences. Yelp users use online check-ins to keep a track of businesses they visit. These check-ins are visible to their friends. This way friends get updated about the users whereabouts and this may have a potential influence on friends future choices.

1.2 Significant Changes in User Behaviour

1.2.1 Identification of Changes

User check-ins is the best way to track customer visits to a particular business. Check-ins after and before a certain date can be compared in order to decide whether the

number of checkins have increased or decreased after that day. The difference of check-ins after and before a review can be taken as a measurement to identify these significant changes in user behaviour. when the check-in difference for a particular day takes a relatively larger value than the others, the number of user check-ins will have a notable increment after that day.

1.2.2 Predict future changes

In literature sentimental analyzing is often used to classify the reviews whether they are positive or negative. Another approach is to classify reviews according to the ratings given by the reviewer. For higher ratings it takes the review as a positive one. The opposite is applied when there is a lower rating. In this paper the dates will be categorized according to the user behaviour after the review. The dates that are probable to expect a sudden increment of user check-ins can be labeled as ‘Increment’. A date that outputs a decrement of user check-ins in a large amount can be labeled as ‘Decrement’. Dates that do not belong to each of the above categories will be considered as ‘Neutral', dates that expected no significant change in user check-ins.

1.3 Trends in Customer Behaviour

1.3.1 Customer Trends Detection

To investigate trends in customer behaviour based on reviews, user check-ins over time can be considered. This approach is more suitable for businesses which have a considerable amount of user check-ins records for a long time as the trends are analysed over time. User behaviour trends can be viewed when the user check-ins are plotted against time. The graph can be divided into sections that the behaviour of users are alike in each section. One section symbolizes the existing period of a trend. Points that divide the sections represent the decay of the existing trend and persist of new trends.

1.3.2 Reasons Behind Customer Trends

Factors that trigger a trend always emerge from the customers. Trends appear because of a common interest among a group of customers. More popularity this trend gets, more customers start to talk about the trend and start to follow the trend.

Communication is the key to spread these trends. User reviews are the method of communication in the Yelp network. If there is a trend, the reviewer should be talking about it. The factors that cause a trend should be the trending topics among reviewers. Common topics in the reviews during a trending period can answer the question of ‘why did this trend appear?’. So the reasons to persist trends can be derived from analysing the content of reviews published during that trending period.

1.4 Problem Statement

Given a dataset including business details, user details, user reviews about businesses and check-ins for a particular business, The problem is to find out the impact of reviews on user behaviour. Sentimental analysis on reviews may categorize reviews into negative or positive. But being a negative review really has a negative impact on the customer behavior? Can a positive review alone increase the number of check-ins for a business? The impact may be dependent on the user profile. What are the factors of a user profile that can lead to a revolutionary feedback? Also some trends are becoming popular from time to time. Can we identify these trends and trends setters using reviews? Can we identify the reasons that persist and decay these trends? Above problems will be addressed in this research.

1.5 Motivation

Business owners are keen to learn about their customer base. Customer churn rate is a major problem that they face. It would be very helpful if the business can identify a pattern which leads to customer churn early and the reasons for that. Not only the customer churn, it is very useful to forecast the increment of customer visits. So that business owners can get ready to supply for the increasing customer demand in the coming future. Businesses can plan their campaigns and marketing strategies according to the predictions. These are the benefits to identify the trend setters at an early stage. By identifying the reviews that can increase the demand for the business it helps to process filtering reviews. Those identified good reviews can be displayed at the top of the reviews list on the web site. This research is useful to such business owners who are interested to know about their future and current customer behaviour. Also the shopkeepers who are willing to know certain patterns that have a negative or a positive effect on the current business are benefited from the research.

1.6 Objectives

General objective: Predicting customer behaviour after a certain date using online reviews and investigate the underlying causes behind trends in customer usage patterns

Specific objectives:

- Study the relationship between customer check visits and online reviews
- Identify review features that leads to changes in customer visits
- Construct data set using features of review text, reviewer, business and check visits
- Annotate the constructed dataset based on customer visit difference before and after a selected date
- Categorize records that can cause negative, positive and neutral effect on business after a given date
- Identify trends in usage patterns and study the relationship between these trends and online reviews
- Identify the reasons to persist the trends in customer behaviour through reviews

1.7 Research Scope

- Businesses which are operating normally will be considered. Businesses which are closed or suspended will not be considered.
- Assume that the geometrical location of a business is not changing.
- Assume that all the reviews are genuine. There are no fake and paid reviews.
- Assume that all the review profiles are genuine. There are no fake profiles and paid profiles.
- Only the restaurant data will be considered.
- Assume that the effect of a review is immediate. Visits within a month will be considered as affected visits.

2. LITERATURE REVIEW

2.1 Impact of customer reviews

2.1.1 Effect on Business Survival

Shopkeepers always worry about the future of their business. They use many marketing and sales strategies to ensure their establishment. Lian, Jianxun, et al. [1] discuss the long term survival of a business in the industry. Further it investigates on how long the business is going to be run or when it is going to be closed according to the user feedback. The purpose of this research is to predict that a business is going to be closed before a certain date.

The authors use data from four different aspects. They are geography, user mobility, user rating and review text. Geographic features consist of location details, competitive shops near by and environmental details. User mobility features support geographical features. Customer satisfaction is measured by ratings and review text. In the training set it includes the data from both open shops and closed shops. Before starting any other work authors have conducted a statistical analysis on the data

For geometrical data predictions spatial metrics were used with the predictors Density, Neighbor Entropy, Competitiveness, Quality by Jensen and Category Demand. Binary classification has been conducted on above predictors using logistic regression. Performance has been evaluated using area under the curve (AUC). Combination of these features gave better results than using them individually. User mobility check-ins were used with predictors Transition Density, Incoming Flow, Transition Quality, Peer Popularity. Even though Peer Popularity gives better AUC performance, combinations of the predictors are more useful. Ratings were given for consumption levels, taste, environ-ment, service quality and overall rating, maximum, minimum and average values were used as features. User reviews were analyzed using text mining techniques and several words were generated which are mostly used for already closed shops and currently open shops separately. Finally experiments were done according to logistic regression(LR), gradient boosted decision tree (GBDT), and support vector machine (SVM). GBDT has given the best performance.

2.1.2 Effect on Business Revenue

The paper by Luca, Michael [2] is all about how customer reviews affect the business revenue. The paper uses review data of restaurants from yelp.com combined with each restaurant's revenue data from Washington State Department of Revenue. First it shows that the ratings are correlated with changes in revenue. There are two identification strategies used in this paper. A regression discontinuity approach was implemented to support the hypothesis “Yelp has a causal impact”. Then fixed effects regression was applied to estimate the heterogeneous effects of Yelp ratings.

The paper has stated the following three core findings.

- Increment of stars in rating leads to increase in revenue in considerable amount
- Rating mostly affected in individual restaurants. Chain affiliations are not much affected.
- Online reviews are a substitute for the reputation of a restaurant.

In addition,

- Readers are not considering all the visible information and responsive to quality changes
- Consumers respond to ratings with information, number of reviews and “elite” reviewers more.
- Reviewers’ Yelp friends network has not much effect on consumers' responses.

2.1.3 Effect on Customer Selection

As the title describes in Gunden and Nefike’s paper [6] the objective of the research is to find out how online reviews affects consumers selection of restaurants. Resturant’s number of online reviews, overall ratings, food quality, service, atmosphere and price are taken as factors that has a high impact on customer decision. Results show that food quality and overall ratings play a major role when selecting a restaurant. A quantitative approach is used in this paper to achieve its results. Choice-based conjoint (CBC) analysis is performed with 353 respondents who check online reviews before visiting a restaurant to investigate what are the most important factors for their selection.

2.1.4 Effect on Customer Ratings

Reviews are unstructured text and rating of the review reflects the attitude of the reviewer. Predicting the rating of a particular review is a popular problem in machine learning. In paper [33] by Asghar, Nabiha tries to address this problem as a multi class classification problem. Class labels are ratings from 1 to 5. The features that derived from review text for the classification problem are unigrams, bigrams, trigramsand and Latent Semantic Indexing. Four models have been compared and the best performing model has been selected. The models are logistic regression, Naive Bayes classification, perceptrons, and linear support vector classification. The data set has been divided into proportion 4:1 for training and test data. 3-fold cross validation was used and Root mean squared error(RMSE) and accuracy were used as the evaluation metrics. Logistic regression has given better results when compared to other models.

The Paper [28] by Farhan, Wae is trying to predict the star rating of the review using factors using features of a restaurant together with months of the review and state. Multiple models have been tried to predict ratings. Naive Bayes, Neural Network, Random Forest and Linear Regression were used. Linear Regression has been identified as the best model. The paper concludes that predicting reviews using features business is a difficult task and hard to increase the accuracy. The reason behind that is personal preference is different from person to person. Also the restaurants change their attributes from time to time.

But paper [29] by Bakhshi, Saeideh, Partha Kanuparthy, and Eric Gilbert has more interesting findings. The paper studies both endogenous and exogenous factors such a s restaurant attributes, local demographics and local weather conditions at the date of the visit. It shows that both endogenous and exogenous factors are affecting customer recommendations. These online effects can be explained using offline theories of experimental psychology. Customer recommendations were viewed through a number of reviews and ratings. Number of reviews were predicted using negative binomial regression and ordered logistic regression was used to model user ratings. When talking about exogenous factors like population, region, education, racial diversity and weather is a concern. On rainy days the ratings like to be negative. This implies the findings by previous studies that sunshine has a positive effect on mood while cloudy weather has a negative effect. Same goes with cold and snow days.

2.1.5 Usefulness of Reviews

Paper [35] by Shen, Ruhui, et al discuss how useful these online reviews are by using the Yelp dataset. The authors argue that the usefulness of the review can be predicted by the votes received by that review. Further since it takes a while to receive votes so it's important to predict the usefulness of the review in early stages. The features that are used in this research are bag-of-words, linguistic, geographical, statistical, popularity and other qualitative features extracted from user, business and review. But paper [36] written by Liu, Xinyue, Michel Schoemaker, and Nan Zhang stated in Yelp itself determining the display of the order of reviews and votes received by the review does not influence the order of viewing reviews in Yelp web site. Which concludes that votes are not the only indicator for a useful review. The paper focuses on the features that can be derived by the words in review text. Finally they propose 10 most indicative words for a useful review and they are day, show, over, star, thing, free, look, chicken, say, lunch.

2.2 Perception of Customer Reviews

2.2.1 Familiarity with the platform

Lim, Young-shin, and Brandon Van Der Heide in their paper [4] investigate how the factors of source, receiver and message text influence the perception credibility of an online review and perceptions attitudes toward the reviewed object. It further studies the effects of review valence, the reviewer profile, whether the receiver is familiar with a customer review site (user) or unfamiliar with customer review sites (nonuser). 241 college students have participated in this research. First they were given reviews to read and then a questionnaire was given which can measure the perceived credibility of the reviews. Finally asked a question that they are familiar with a customer review platform to identify the participant is a user or a nonuser. Results states that The receivers unfamiliar with the platform were not influenced by external factors like number of friends, the number of reviews, etc. But for the receivers who are familiar with the platform number of friends and the number of reviews were affected by the attitude towards the reviewed object.

2.2.2 Learning rate of the users

Anenberg, Elliot, Chun Kuang, and Edward Kung in their paper [27] show that customers learn about the quality of the restaurants by online reviews. Thus the demand for certain restaurants depends on how good these reviews are. Further the paper states that the learning rate from online reviews are higher around city centers and as well as places where there are educated people. So the quality of the restaurants are higher in these areas. The main thing to extract from this paper is that there is a relationship between geographical area a restaurant is situated in and online reviews and it affects the quality of the restaurant as well as customer behaviour.

2.2.3 Cultural influences

The paper [30] by Nakayama, Makoto, and Yun Wan compares the reviews between western customers and japanese customers towards Japanese ethnic restaurants. The researchers are experimenting about how the customer's cultural background plays an important role when perception of online reviews. The study focuses on four main aspects of a restaurant: food quality, service, ambiance and price fairness. Price fairness is highly considered when giving a positive comment and high ratings while ambiance is a key factor for unfavorable reviews in Japanese society. In western culture service is the most important feature for both negative and positive responses.

Author Kong, Angela, Vivian Nguyen, and Catherina Xu in Paper [34] are trying to identify the features that affect the higher restaurant ratings. Restaurants in the US, UK, Canada, and Germany were considered for the study. Six main features have been identified which are mutual to all countries. They are availability of street parking, ability to make reservations, review count, casual ambience, noise level, and attire. In the US, restaurants with divey ambience get higher ratings. Also Americans are more vocal, negative, and service oriented in their reviews. In US and Canada ratings for restaurants with parking facilities are high as the number of drivers are large. In Europe where public transport is more popular parking does not gain favourable ratings. But restaurants that serve alcohol get higher ratings as the age limit for drinking is lower.

2.2.4 Customer trends

Paper [28] by Farhan, Wae is giving insights on how well the restaurants are performing based on the features of the restaurant and give clues to improve the customer satisfaction. The analysis is done comparing the average ratings given to the relevant restaurant. Restaurants which are good for kids are highly rated if they serve happy hour meals. Also ratings for restaurants that do not have TVs, serve alcohol and allow smoking. There's an annual trend of having more ratings during the very beginning of summer and restaurants which are serving alcohol are popular in winter over summer. Also Ratings are higher in the month of February as there is valentine day and people tend to visit restaurants. In the Paper [29] by Bakhshi, Saeideh, Partha Kanuparthy, and Eric Gilbert shows that the number of reviews shows an increment during summer and Fall.

2.3 Features of Online Reviews

This section summarizes the features online reviews affecting the customer behaviour identified in various research papers. These attributes can be divided into three categories; features of business, review and reviewer. It seems like features of restaurants and reviews are studied but features of reviewers are not studied deeply.

Table 2.3.1 Features of a Restaurant

Category	Features
Quality and Reputation	• customer ratings • number of reviews • reviews arrival rate • Number of checkins • monetary • advertising
Services	• Food quality • type of alcohol served • noise level • price range • happy hour • smoking options • has tv • wi-fi • good for groups • good for dancing • food cuisine

	• carryout
Location	• latitude, longitude • state, city, town • median income • population density • diversity index • higher education • number of shops • check-ins of nearby shops

Table 2.3.2 Features of a Review

Category	Features
Text	• unigram • bigram • Trigrams • TF-IDF • Latent Semantic Indexing
Attitude	• review valence • review text • polarity • subjectivity
Review date	• month • mean temperature • precipitation • snow

Table 2.3.3 Features of a Reviewer

Category	Features
Profile reputation	• number of friends • number of reviews • Avg and variance of the votes received by

2.4 Predicting User Behaviour using Classification

2.4.1 Classification

Classification is a data analysis process that is intended to find a generic model to

group various data instances to the right category they belong to. Basically here the classification algorithm identifies the most suitable category for a new data instance based on a training data set. Training data set is a collection of data instances which are grouped into classes. The number of classes is based on the problem.

If it is a medical diagnosis where we have to classify the report as negative or positive; it is called a binary classification because there are only 2 categories. If there are more than two categories the problem becomes multiple class classifications. If we want we can solve multiple class problems in many ways. First one is to transform the problem into several binary classification problems. We can train a single classifier per class, with the samples of that class as positive samples and all other samples as negatives. Or we can deploy pairwise classifiers and finally combine the output.

There are many classification algorithms. Some of the most popular classification algorithms are listed below. Most of them can be used for binary classifications as well as multiple class problems.

- support vector machines
- Bayesian network
- nearest neighbor
- decision tree
- Random Forest
- AdaBoost
- Artificial neural networks.

2.4.2 Imbalanced Data

Number of instances from each class matters. If it is a balanced dataset, the number of instances from each class is nearly equal. If the number of instances from a certain class is greater than others, the dataset is said to be imbalanced. Most of the time imbalanced data sets are highly available over balanced data sets. Medical Diagnosis, credit card fraud detection and Network Intrusion Detection are some of the areas that have such data. In literature it states that these classification algorithms are performing well on balanced datasets. When it comes to imbalanced datasets low accuracy can be observed. This is due to skewed distribution of class instances. The classification algorithm tends

to be biased towards the majority class. Because of this minority class rules will not be captured during the training phase[11]. So there is a higher chance of misclassifying the rare class instances which leads to poor performance.

There are two main approaches to make data sets balanced.

1. Oversampling - increase the rare class instances
2. Undersampling - decrease the instances from majority class

One of the problems that can arise with data level approaches is losing information. Another thing is that showing a false proportion during and this can be affected on test data in future. But these methods can not be completely rejected as they are widely used in the industry and show good results. But it's better to be cautious when using these techniques. Another way is to follow an algorithmic approach. Here we can choose an appropriate inductive bias.

Most of the outlier detection and anomaly detection problems fall into imbalanced data problems. In another perspective, a data set which has a lot of noise also can be viewed as imbalanced data problems. Both of th

2.4.3 Binary Class Classification on Imbalanced Data

There are many researches conducted on binary class classification on imbalanced datasets. They have followed different types of approaches.

2.4.3.1 Converting into Balanced Data

Kuncheva, Ludmila I., et al in their paper[10] states that the most common method to solve imbalanced data classification is equalising the number of class labels and then training the classifier on newly obtained balanced data sets. Geometric mean (GM) is used to measure the accuracy of the model. The paper is trying to fill the gap between the practical approaches and theoretical proof behind these approaches. Basically the authors prove that GM can be improved by instance selection methods and theoretical explanation is given for such improvement. For instance selection methods paper only consider undersampling methods.

To verify the theoretical concepts, an experiment has been carried out for 12 instance selection methods on 66 data sets. Results show that these selection methods can be improved further. The 12 instance selection methods includes Random undersampling (RUS), ensemble of RUS combined by majority vote, an AdaBoost variant that performs both re-weighting, RUS in each itera-tion, Evolutionary undersampling (EUS), AdaBoost-like ensemble of EUS, Particle-swarm optimisation, Tomek links(TL), condensing followed byTL and Neighbourhood cleaning rule.

There are two main findings of the research paper.

- “GM is non-monotonic with respect to the number of retained instances,which discourages systematic instance selection.”
- “Balancing the distribution frequencies is inferior to a direct maximisation of GM.”

2.4.3.2 Algorithmic Approches

2.4.3.2.1 Neural Network

Paper by Wang, Shoujin, et al [11] studies the performance of deep neural networks binary classification on imbalanced data sets. In deep learning, the most common loss function is mean squared error (MSE). It captures all the errors from the full data set and then calculates their average. When the data is imbalanced the majority class contributes to loss values more and the loss function is biased towards the majority class.

In order to solve this problem two new loss functions have been introduced in the paper for training so the learning algorithm gets more sensitive to the minority class. They are called mean false error (MFE) and mean squared false error (MSFE). Here the loss functions calculate the averages of error from each class separately and add them together. With this approach the authors were able to capture the errors from both classes equally. The novel loss functions have been tested on 8 datasets and 4 network structures. The percentage of the minority classes of these data sets were 20%, 10% and 5%. The performance was measured using F-measure and AUC and MSE was used as the baseline. The results show that novel loss functions outperform the conventional loss function MSE.

2.3.3.2.2 Deep Reinforcement Learning

The paper of Lin, Enlu, Qiong Chen, and Xiaoming[12] introduces a general imbalanced classification model based on deep reinforcement learning for binary classification. The classification problem was transferred into a sequential decision making process which was later solved by a deep Q-learning network. The agent classifies instances from the dataset one by one and earns a reward for the action. To make the agent more sensitive to the minority class reward for classifying has made large. Finally an optimal classification policy will be output by the agent with the help of reward function and learning environment.

As the baseline a deep neural network trained with cross entropy loss function was used. The algorithm was compared with the other approaches found in the literature to handle imbalanced data. Methods are oversampling minority classes, undersampling majority classes, using mean false error as the loss function, assigns higher misclassification cost to minority class and low cost to majority class in loss function. Finally adjust the model decision threshold in test time by incorporating the class prior probability. To evaluate the performance G-mean and F-measure have been used. Experiment was conducted with several datasets imbalanced ratio from 0.05% to 10%. Results show that the proposed method outperforms other methods for both balanced and imbalanced data sets.

2.4.4 Multi Class Imbalanced Classification

The papers on multiclass imbalance classification are rare. One of the main reasons is that the multiclass problem can be converted into binary classification and the imbalanced data is converted into balanced data. Most research done on binary imbalanced classification by using algorithmic approach has suggested applying the suggested methods to apply on multiclass problems in future[11]. So this is a novel research area to explore.

2.5 Exploring Trends in Customer Behaviour

2.5.1 Time Series Data Mining

A time series is a set of observations on the values that a variable takes at different times. It is a series of data points which are ordered according to time. In addition to that time series is the most common representation of temporal data. Patterns and trends over time can be discovered by analysing time series data. Data mining is the process of discovering patterns in large data sets by using statistical and machine learning techniques. Most data mining problems consist of temporal data. Hence exploring behaviour in time series has become an emerging field in data mining and the concept is called time series data mining. In literature Major time series related tasks are query by content , anomaly detection, motif discovery, prediction, clustering, classification and segmentation [15].

2.5.2 Time Series Segmentation

2.5.2.1 Identify Trends Using Time Series Segmentation

Time series segmentation can be used to identify the trends in customer behaviour over time. Time series is considered as a sequence of discrete segments of finite length. Each segment is a phase or a state. Each phase or state is generated by a system with distinct parameters. External events or internal systematic changes cause changes in the behaviour of data distribution. Data points that do the partition of these phases are called critical points or change points.

2.5.2.2 Identify Segments in Timeseris

There are two approaches to identify segments in time series. One is to identify the change points. The other one is inferring the most probable segment locations and the system parameters that describe them. While the first approach tends to only look for changes in a short window of time, the second approach takes the whole time series when deciding which label to assign to a given point.

The segmentation problem can be expressed in three ways [18]

- Given a time series T, produce the best representation using only K segments
- Given a time series T, produce the best representation such that the maximum error for any segment does not exceed some user-specified threshold, max_error
- Given a time series T, produce the best representation such that the combined error of all segments is less than some user-specified threshold, total_max_error.

2.5.2.3 Representation of Time Series Data

The most common representation of time series segments is piecewise linear representation. One of the main techniques used to solve time series segmentation is Piecewise Linear Approximation [16]. In addition to that, the most used approximation is Piecewise Linear Interpolation. In mathematical context, linear interpolation is a curve fitting method which uses linear polynomials to construct new data points within the range of a discrete set of known data points.

There are 3 main approaches used in linear interpolation.

- Sliding windows - Segment is grown until it exceeds some error threshold
- Top-down - Recursive partition of time series until some stopping criterion is met
- Bottom-up - Segments are iteratively merged starting from most suitable approximation

2.5.2.4 Piecewise Linear Approximation

In paper [17] authors Keogh, Eamonn, et al. conducted a review on segmenting time series data in the context of piecewise linear approximation. A comparison is done for most of the existing methods and show the major failures of them in a data mining perspective. So they propose a novel algorithm which surpasses the performance of other rest. As the paper states there are two approaches that can be used to piecewise linear approximation. They are linear interpolation and linear regression. Linear interpolation gives a smooth curve while linear regression produces disjoint segments on some data sets.

In order to apply either of approximation, research paper introduces a novel clustering

based method named ‘SWAB’ which is a combination of two approaches sliding windows and bottom up. This new method reduces the disadvantages from sliding windows being an online segmentation technique while retaining the advanced qualities of bottom up approach.

2.5.3 Change Point Detection

A sudden or unexpected change in data over time can be identified as a change point. There are many applications to detect these abrupt points in time series data. Also there are many applications that use these techniques such as medical condition monitoring. Change point detection can be treated as a binary classification problem. All the state transitions can be labeled as change points and all the other within state points labeled as a none change point.

If there are labeled data supervised methods like decision trees, naive Bayes, support vector machine (SVM), nearest neighbor, hidden Markov model can be used. The input features for the training should be sources of the change and the number of classes will be the number of states in the time series data.

Unsupervised learning methods have to be used if there is no labeled data. There are many unsupervised approaches used to detect change points that can be categorized as follows.

- likelihood ratio methods
- subspace model methods
- probabilistic methods
- graph based methods
- clustering methods

Arif and Siti Nur Afiqah Mohd in their paper [37] change point analysis has used to find significant changes in meteorological data rainfall, temperature and humidity. Study has used a method with two approaches namely CUSUM and bootstrap. CUSUM was used to find the patterns and trend in time series data and bootstrap was used to find the occurrences of abruct points. According to the paper there are many papers that have used the same approach. This approach originated from the article written by Taylor,

Wayne A in 2010 [38]. Kass-Hout, and Taha in their paper [39] tried to monitor the changes in visits for the emergency department due to influenza-like illness. There they evaluated the existing change point analysing methods to predict the incidents. CUSUM analysis was compared with the structural change model and Bayesian method. CUSUM analysis method was more appropriate and accurate when compared with the other two.

2.5.4 Perceptually Important Point Detection

Apart from change point detection, detecting perceptually important points is a similar concept. Perceptually important points (PIP) represent the minimal set of data points which are necessary to form a pattern. This approach has been used in many research papers. Most of the papers follow the algorithm in figure 2.5.4.1. Here we have to define a number of segments. When there are multiple graphs this approach might not be the ideal one as different graphs have different numbers of segments.

Fu, Tak-chung, et al. [21] propose a new time series representation according to the importance of data points. The graph is represented by a specialized binary tree which is constructed by the data based on the data point importance. The important list is created by identifying PIPs and calculating their importance. The cutting points which separate the segments are derived from the binary tree itself. This approach can be used in many time series mining problems. Also the algorithm can be improved further by customizing for certain domains. Paper was extended by the authors later by favouring unique behaviours in financial time series data [22].

Chung, Fu-Lai, et al. in their paper [19] time series segmentation was viewed as an optimization problem and they proposed evolutionary computing techniques to solve the problem. The main objective of this problem is to identify the PIPs among a set of time point templates. As the fitness evaluation function direct point to point distance is used. It measures the similarity between two time series segments. However later authors Yu, Jingwen, et al. [20] argue that this is not an appropriate method to calculate the similarity of two sequences with different length. Further it investigates the limitations of the above method and proposes a new evaluation function based on pattern distance. The new method calculates the dissimilarity of two time sequences and this is a measure of similarity of trends.

Figure 2.5.4.1

Pseudo code of the PIP identification process

```
Function Pointlocate (P,Q)
   Input: sequence P[1..m], length of Q[1..n]
   Output: pattern SP[1..n]
Begin
   Set sp[1]=p[1], sp[n]=p[m]
   Repeat until sp[1..n] all filled
   Begin
      Select point p[j] with maximum distance to the
        adjacent points in SP (sp[1] and sp[n]
        initially)
      Add p[j] TO SP
   End
   Return SP
End
```

2.6 Extract Trending Topics in Reviews

2.6.1 Structured Version of Text

Text is unstructured data. It is needed to be preprocessed before applying data mining techniques like text mining, natural language processing and information retrieval. A structured version of data is output after going through the preprocessing step. This structured data can be divided into two groups: single word and multiple words. Single word is called a bag of words. The main difference between a bag of words and multiple words is that multiple words preserve the relationship between words so that the semantic meaning of the sentence is maintained.

2.6.2 Frequent Itemset Mining vs Association Rule Mining

Frequent patterns appear frequently in a dataset. Frequent itemset is made up from one of these patterns. Frequent itemset mining finds interesting frequent itemsets that exceed minimum threshold value among transaction data. This threshold value is a measure that indicates the minimum number of items that contains. The idea of frequent pattern mining was involved with transactions where market basket analysis was taken as the most popular example. Here products that are sold together are analyzed. In simple, Frequent itemset mining shows which items appear together in a transaction or relation.

Association rule mining is about finding the relationships among these items. In view of market basket analysis, what is the probability of buying a certain product after purchasing two correlated products. However both concepts go together when applying to problems.

2.6.3 Support and Confidence

There are two main measurements when it comes to frequent itemset mining. They are Support and Confidence. These two measures evaluate how interesting the patterns are. Both measures indicate how frequent items appear in transactions. Support is a measure of absolute frequency and confidence is a relative frequency. Support indicates how many times the items appear together out of all the transactions. Confidence indicates how many transactions follow a certain rule. Basically support is a measurement of frequent items that does not consider relations between them and confidence is a measurement of association rules among items. Following are the notations.

$$\text{Support}(A \rightarrow B) = \text{Support\_count}(A \cup B)$$

$$\text{Confidence}(A \rightarrow B) = \text{Support\_count}(A \cup B) / \text{Support\_count}(A)$$

2.6.4 Frequent Itemset Mining in Text

A sentence consists of words. There can be common words among multiple sentences. Similarly paragh consists of words and multiple paragraphs have frequent word sets. These common words can be viewed as frequent item sets. Words will be an item in the item set. So the frequent pattern can be formed by multiple words. Mining these patterns will give a basic idea about the common topics in text rather than extracting single keywords.

Frequent itemset mining mainly applied on structured data uses many algorithms. Apriori, FP-Growth, Eclat are some of the widely used algorithms. Maylawati, D.S. [26] states that only a small percent of frequent itemset mining algorithms are applied with text data. There are many more algorithms that have not been implemented in text.

This opens a new area of research to text mining, information retrieval, and natural language processing

2.6.4.1 Text Summarization Using Frequent Pattern Mining

In the paper by Hu, Minqing, and Bing Liu [24] the main objective is to generate feature based summary of a product from a large number of reviews. With the increasing number of reviews neither users nor product owners can't get a clear overall idea about products. So providing a summary is very helpful to both parties. The significant difference from other text summarization is features of the products will be considered and the opinion regarding each feature is positive or negative. The research is conducted by using data mining and natural language processing techniques. To extract the features association rule mining approach is used which is based on Apriori algorithm. The minimum support is 1% and several pruning methods have been applied.

Most people share traffic related experiences online social media. Analysing traffic events from social media helps drivers to adopt changing traffic conditions as well as government authorities to manage traffic conditions in future. Most of the traffic related information is based on a list of single keywords. This leads to lots of noisy data and false information. The main objective of the research by Xu, Shishuo, et al.[25] is to filter noisy data so that summarizing information represents real world traffic. To filter only positive data containing messages, First messages were queried using a predefined set of single keywords and extract matching messages. Keywords included words related to traffic data like Accident, Traffic, Blocked, etc. The association rules between words in filtered messages were mined by Apriori algorithm. A set of new words were introduced to retrieve future queries. Further it classifies the traffic events into five categories by supervised machine learning methods.

3. METHODOLOGY

3.1 Experiment

The main objective of this experiment is to find the impact of online reviews on user behaviour. Lian and Jianxun have taken user online check-ins as an indicator of the customer transitions among businesses to trace user mobility [1]. As evidenced by their paper, customer behaviour can be tracked through the customer online check-ins.

The study carried out two main approaches to achieve the main objective. One is to predict changes in customer behaviour after a certain date based on the reviews received by the individual business. Further to identify customer visits increases or decreases than the average or remain the same after that day. This date can be referred to as the prediction date. The other approach is to identify customer trends and trending topics for each business. As the initial step a statistical analysis was conducted to get a better understanding about the data set. There the relationship between customer online activities and reviews was studied.

3.2 Data

A dataset from a famous Yelp challenge was used. The data set was published for 2019 (round 13) and it is publicly available for research purposes. Data set consists of the following json files in Table 3.2.1. Figure 3.2.2 is a diagram of relationships among the above entities.

Table 3.2.1

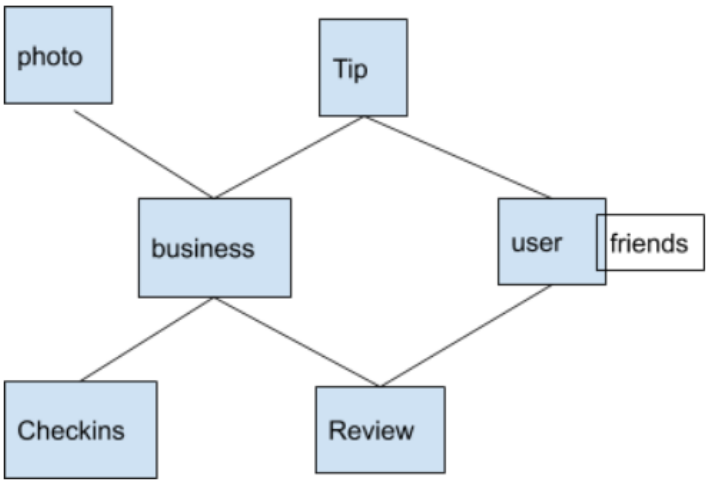
Yelp data set

File name	Description	Attributes
Business.json	Details of a businesses	Business_id, name, address, city ,state, postal code, latitude, longitude, stars, review_count, is_open, attributes (RestaurantsTakeOut, BusinessParking), categories

		(Mexican, Burgers, Gastropubs), opening_hours
Review.json	Details of online reviews posted by users	Review_id ,user_id, business_id, stars, date, text, useful, funny, cool
User.json	Details of registered users in Yelp.com	User_id, name, review_count, yelping_since, friends, useful, funny, cool, fans, elite, average_stars, compliment_hot, compliment_more, compliment_profile, compliment_cute, compliment_list, compliment_note, compliment_plain, compliment_cool, compliment_funny, compliment_writer, compliment_photos
Check-in.json	Details of user visits at a business	Business_id, date
Tip.json	Shorter than a review. Contains an important fact.	Text, date, compliment_count, business_id, user_id
Photos.json	Details of photos	Photo_id, business_id, caption, label

Figure 3.2.2

Relationships among Yelp Data set entities



3.3 Understanding the Data Set

In this section it presents the results of the data analysis performed on the yelp dataset. Table 3.3.1 shows a fundamental statistical summary for each entity. Data was extracted from the json files that were provided by Yelp and stored in a relational database for easy access.

Table 3.3.1

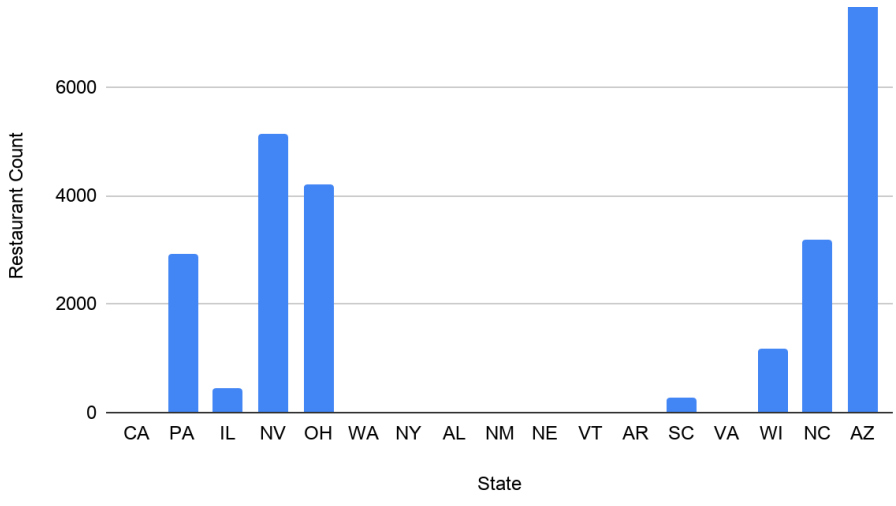
Basis stats of the dataset

Entities	Count
Reviews	5,228,029
Users	1,609,073
Check-ins	1,950,474
Businesses	158,525
Restaurants	42,296

The dataset has various types of businesses. According to Table 3.3.1 most of the businesses are restaurants. To filter out the restaurants, businesses which have ‘Restaurant’ in the categories field were selected. When comparing the state wise restaurant counts in the USA according to the dataset, state ‘Arizona’ has the most number of restaurants. (Figure 3.3.1)

Figure 3.3.1

Number of restaurants in each state in the USA



When going through the data set not every business has a considerable amount of check-ins and review records. They might be newly opened or not so popular businesses. According to Figure 3.3.2 Some businesses have a good amount of check-ins but not reviews. Most of the time check-ins are higher than review count. Yelp uses the concept of ‘Tips’. Tip is a short comment but provides an important fact about the business by using few words. Sometimes customers tend to read and post tips over reviews. But going through the data set a considerable gap is found between review count and tips count. Reviews and tips can be taken as an indication of customer interaction with a business.

Figure 3.3.2

Number of Reviews, Check-ins, and Tips for selected five businesses in 2018

One business was selected and its check-in counts, review counts and tips counts were examined. Following is the details of the selected business.

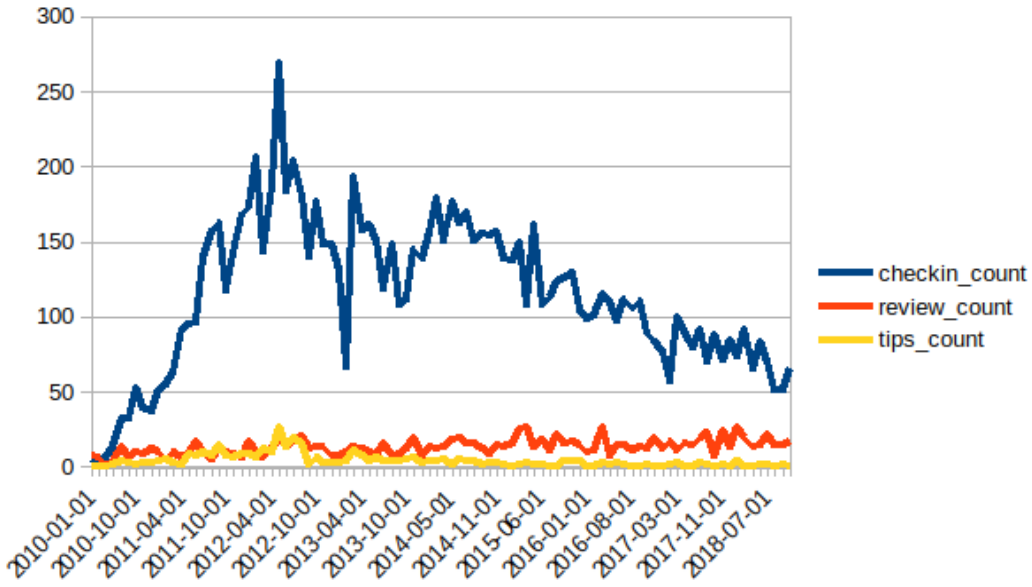
{
  "stars": 3.5,
  "review_count": 1507,
  "attributes": {
        "Alcohol": "u'full_bar'",
        "OutdoorSeating": "True",
        "Ambience": "{'romantic': False, 'intimate': False, 'classy': False, 'hipster': False, 'divey': False, 'touristy': True, 'trendy': False, 'upscale': False, 'casual': False}",
        "NoiseLevel": "u'very_loud'",
        "BikeParking": "True",
        "BusinessParking": "{'garage': True, 'street': False, 'validated': False, 'lot': False, 'valet': False}",
        "WiFi": "'no'",
        "BusinessAcceptsCreditCards": "True",
        "RestaurantsGoodForGroups": "True",
        "GoodForKids": "True",

```
            "RestaurantsReservations": "False",
            "GoodForDancing": "True",
            "HasTV": "True",
            "BestNights": "{'monday': False, 'tuesday': False, 'friday': True, 'wednesday': False, 'thursday': False, 'sunday': True, 'saturday': True}",
            "RestaurantsPriceRange2": "2",
            "Music": "{'dj': False, 'background_music': False, 'no_music': False, 'jukebox': False, 'live': True, 'video': False, 'karaoke': False}"
  },
  "categories": "Festivals, Bars, Nightlife, Arts & Entertainment, Casinos, Shopping, Local Flavor",
}
```


Figure 3.3.3 compares the monthly check-in count, review count and tips count of the selected business from 2010 to 2018. We see that check-ins have increased from 2010 to 2012. After that there's a sudden drop. After 2014 check-ins have started to drop. Meanwhile the review count has been increased little by little from 2010-2018. Tips count remains stable most of the time. But in 2012 there is a sudden rise in tips count too.

Figure 3.3.3

Monthly check-in count, review count and tips count over time for business "Fremont Street Experience"



Next, Relationship between reviews and customer visits was observed. Main objective was to find out whether the customer visits increased with good reviews and whether customer reviews decreased with bad reviews. Figure 3.3.4 and Figure 3.3.5 shows the relationship between reviews counts and customer check-ins according to the dataset. Here we don't see any strong correlation between good review count and customer check-ins. Check-in count is expected to increase with the number of good reviews. In

fact the Covariance is 17.34 and Correlation is 0.36. On the other hand the correlation between bad review count and user check-ins are very weak. The covariance is 2.34589947 and correlation is 0.28. Here also we don't see the expected behaviour where the number of check-in counts decrease with respect to a number of bad review count.

Figure 3.3.6 illustrates check-in count within a month after and before a certain day in year 2018 for a selected business. Here check-in counts are shown in the time series data. Sometimes check-in count has been increased and sometimes check-in count has been decreased after a certain date.

Figure 3.3.4

Relationship between good review count before a certain date and customer check-ins after that date

Figure 3.3.5

Relationship between bad review count before a certain date and customer check-ins after that date

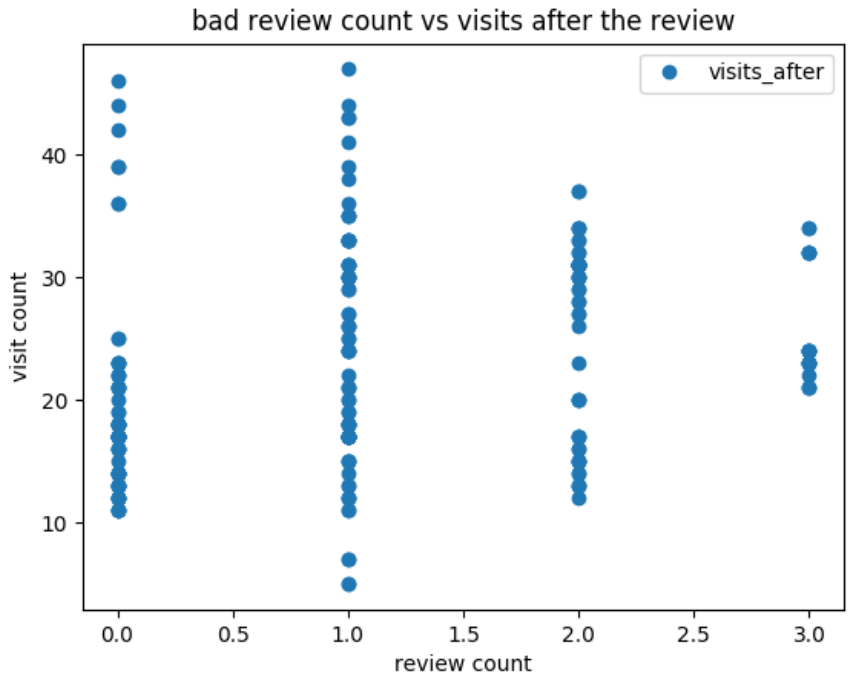
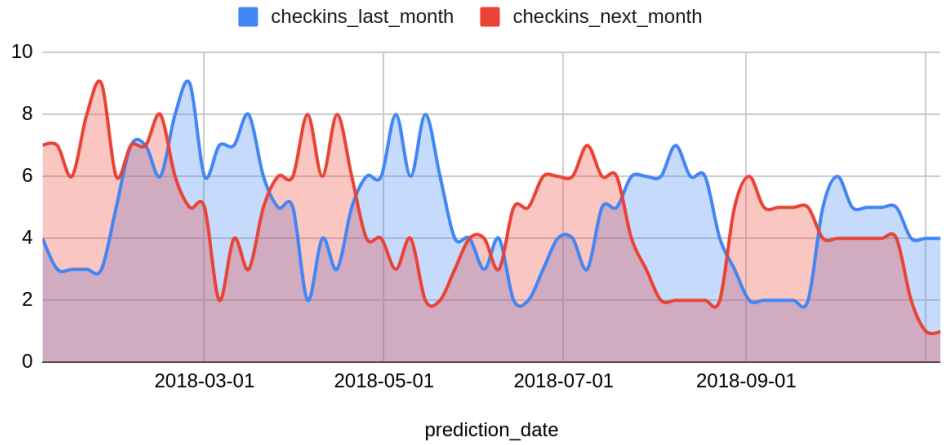


Figure 3.3.6

Monthly check-in count before and after a certain date over the time for a business



The difference of check-in count throughout the year 2018 is shown in Figure 3.3.7. Positive values indicate that next month check-in count is higher than the previous month check-ins. Further it indicates that increased check-in has been increased after

the selected day. Meanwhile, negative values show that check-ins have decreased after that day.

Figure 3.3.8 shows the statistics for the distribution of difference of check-in counts. The graph shows that the check-in difference is between -18 and 20. According to the above statistics the average difference should be around zero and can be between 0 and 13. Values higher than 13 and lesser than 0 show that there is a significant change in customer check-ins after a respective date.

Figure 3.3.7

Difference of check-in counts before and after a certain date over the time for a business

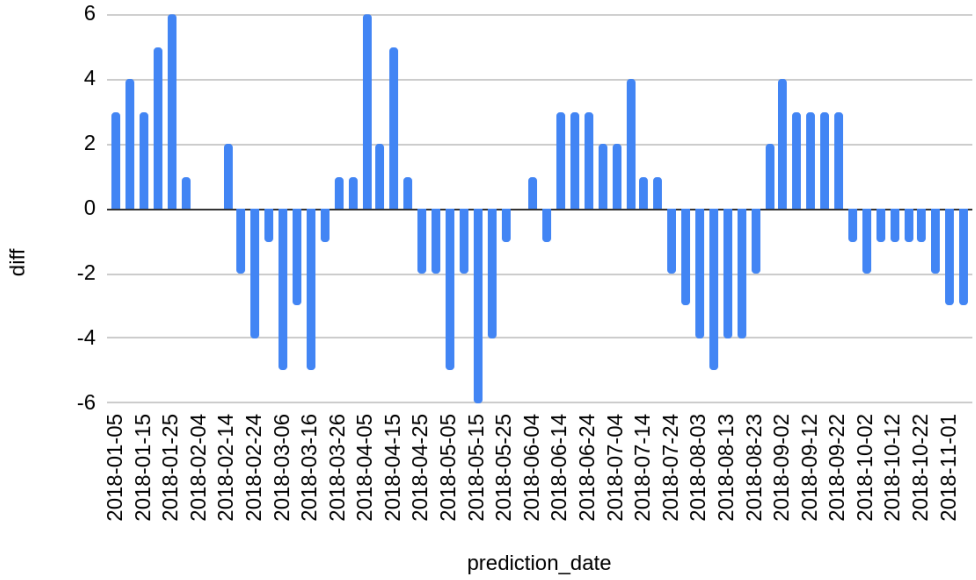
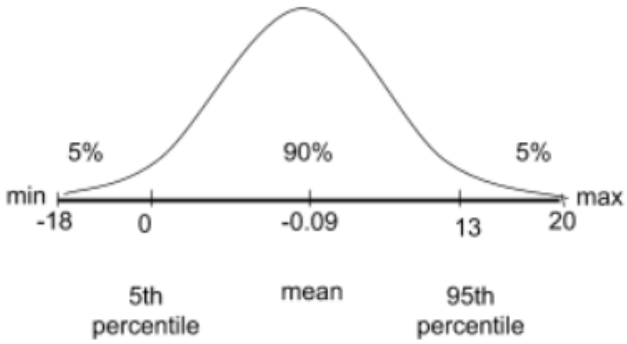


Figure 3.3.8

Distribution of check-in difference



3.4 Predicting User Behaviour

The first step of the study is to predict customer behaviour after a certain date. This date was referred to as the prediction date. So far there is no annotated data set for this requirement. A new data set was generated from the existing data set. New features were derived from attributes of reviewer, business and review text. Annotating the data set was able to be conducted without any human intervention. The process will be automated and it follows guidelines in the next paragraph.

In Order to annotate the data set, dates that have significant impact on customer behaviour needed to be identified. For this check-ins before and after a certain date were compared. Dates that have a remarkable difference between check-in counts were recognized as the dates that can expect a major change in user check-ins afterwards. The remaining dates are the ones which do not expect change in user check-ins. Identified significant changes further grouped into increment or decrement based on the magnitude and sign of the difference of check-in counts when labeling.

A classification approach was followed after creating the training data set. Results from several classification algorithms will be compared to select the best model.

3.4.1 Feature Selection

For a classification problem identifying relevant features is important. Better predictions can be obtained for better features. One of the main objectives of this research is to predict changes in customer visits after a given date and a classification approach was used. Features of review text were considered along with the features of reviewer and business as the variables.

More informative details can be discovered by original data by applying various techniques on the original data set. New features can be derived by converting their data type, aggregating features and changing the data range. Also missing values and highly correlated data should be handled before the classification. Tables in Appendix I, Appendix II and Appendix III shows the new features can be derived from the existing dataset and the naive features that are unique to this research. These features were derived by considering factors studied in the literature which affects customer behaviour

as identified in section 2 (literature review) in this report.

The features can be categorized into three groups. First category is review features which can be retrieved by review date, review text and replies received by the review. The second category is reviewer details which includes the static information about the reviewer's Yelp user account and reviewers historical activities which describes the nature of the user. The third category describes the features of the business. The popularity and the category of the business was measured using these features.

In Order to generate the training set, details related to the previous month were collected for the feature set. These training set features were derived by aggregating the basic features introduced in tables of Appendix I, II and III. There were 42 features in the training set describing reviews, reviewers and business status within the past month before the prediction date. All of these features are numeric. These aggregated features from Reviewer and Reviews are described in tables of Appendix IV and V. Features of the business status are the same as the features in Appendix III. Features derived from the prediction date are week_of_year and day_of_year.

From all the features described above, candidate features for the training set were selected by comparing the correlation between features. Pearson correlation matrix was used in order to generate the values. Figure 3.4.1 shows the correlation matrix between selected features and table 3.4.1 consists of selected features for the training dataset

Table 3.4.1

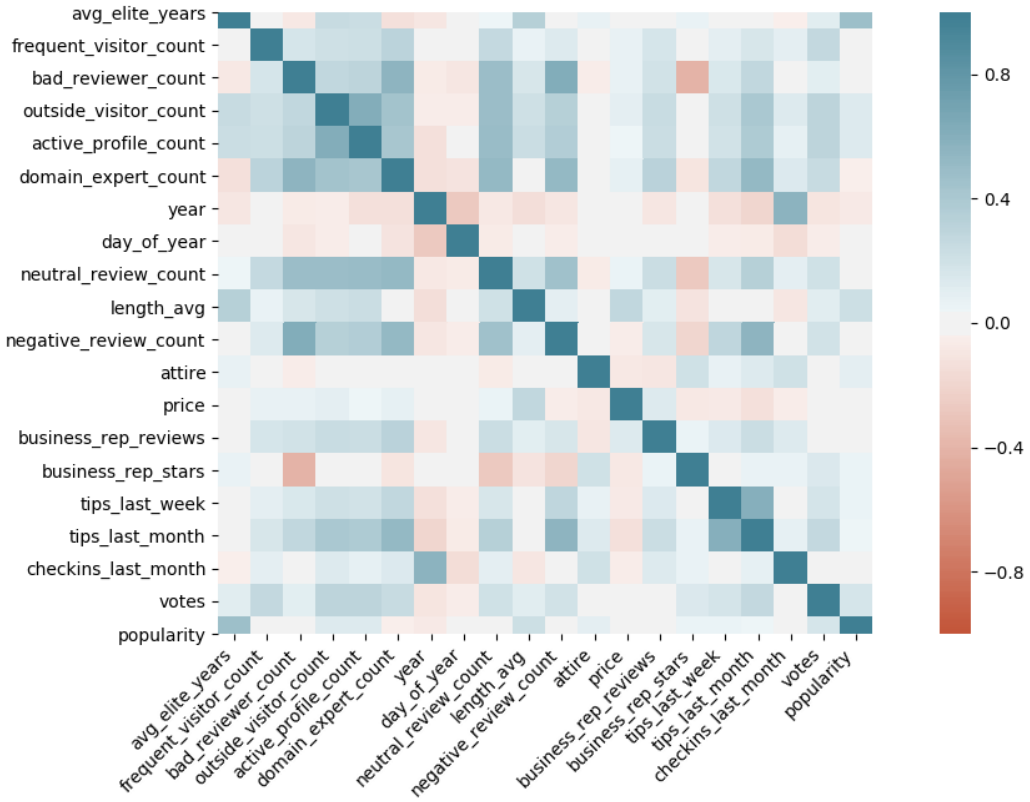
Features of the training set

Features of Reviewer	**Features of the Review**	**Features of Business**	**Features of the Prediction date**
avg_elite_years	length_avg	attire	year
frequent_visitor_count	votes (cool_vote_count + funny_vote_count + useful_vote_count)	price	day_of_year
frequent_visitor_count		business_rep_revi	

		ews	
bad_reviewer_count		business_rep_stars	
outside_visitor_count		tips_last_week	
active_profile_count		tips_last_month	
domain_expert_count		checkins_last_month	
popularity (avg_fans + avg_friends)		negative_review_count	

Figure 3.4.1

Correlation matrix between selected features



3.4.2 Class Labels

Class labels indicate how reviews posted in a certain period before the prediction date affect customer behaviour. If user check-ins are significantly increased after the

prediction date, the class label for that date is ‘Increment’ (+1). Which indicates that the reviews which are posted before that date have a positive impact on user behavior. If user check-ins are significantly decreased after the prediction date, the class label for that date is ‘Decrement’ (-1). Which indicates that the reviews posted before the prediction date have a negative impact on user behavior. If there’s no significant change in user check-ins after a certain date, it is labeled as neutral (0). Which indicates that there is no impact on user behaviour by previously posted reviews.

3.4.3 Annotating the Training Set

The training set consists of features belonging to three categories and three class labels. The three categories are features of review, reviewer and business status. Three class labels are ‘Increment’ (+1), ‘Decrement’ (-1) and ‘Neutral’ (0). As the next step, the data set should be annotated. The most important step of annotating the data set is to identify reviews that belong to each class.

Change in user behaviour in a business is identified by comparing the Check-ins within a month before and after a chosen predicted date. Change in use behaviour or the checkin difference was calculated according to the following equation.

$$\text{Check-in Difference} = \left[\text{Check-ins next month wrt prediction date}\right] - \left[\text{Check-ins last month wrt prediction date}\right]$$

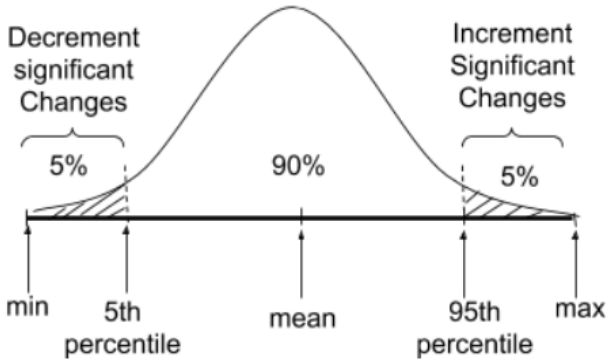
This makes the difference to be negative or positive. Positive value means the Check-ins have increased and the negative value means that the check-in has decreased. If the absolute value of the magnitude takes a very large value, there is a significant change in user check-ins. When considering the distribution of the checkin difference these values are located near the edges as described in Figure 3.4.2.

A lower bound and upper bound was set by looking at the distribution of the difference values to identify significant changes. Values which are greater than the 95th percentile were labeled as ‘Increment’ (+1) which implies reviews before the prediction date have a positive impact on user behaviour. Values lower than the 5th percentile were labeled as ‘Decrement’ (-1) which implies that reviews before the prediction date have a

negative impact on user behaviour. Rest of the records were labeled as ‘Neutral’ (0), which means there is no significant impact from reviews posted before the prediction date.

Figure 3.4.2

Distribution of the check-in different



3.3.4 Training Data Set

To generate the data set restaurants were selected among the business. Restaurants are the most number of businesses that the data set has and In literature most papers are written considering the restaurant industry.

Bakhshi, Saeideh, Partha Kanuparthy, and Eric Gilbert in their paper [29] states that weather conditions directly affect the selection of restaurants by the customers. Since the original data set does not contain the weather details, Restaurants from the same geographical area were selected that one may mitigate the issue. Restaurants located in the state Arizona, USA were selected as it is the state that has the most number of restaurants in the USA according to the data set.

The popularity and the reputation of the business can be represented by review stars and number of reviews [2][3]. The data set has been provided with the most recent detail about the business. So the operation level of the business at a very early date can’t be retrieved from the original data set. Therefore most recent data was used which are Reviews, Tips and Check-ins from 2017 and 2018. Prediction dates were generated from 2017-01-01 to 2018-12-31 with the interval of 5 days.

The training data set contains 4,190 records including 7559 number of businesses. Following is the class label distribution.

Neutral (0) - 1,035 (25%)
Increment (+1) - 1,882 (45%)
Decrement (-1) - 1,272 (30%)

3.3.5 Classification and Model Evaluation

According to the literature review, the most popular algorithms that are used for this data set are RandomForest (RF), XGboost (XGB), Neural Network and Logistic Regression. All the four algorithms were used and the best performing model for the data set was selected. The evaluation metrics that were used were accuracy, precision, recall and F1-score.

The best performing model underwent a hyperparameter tuning process to enhance its performance. Randomized Search with K-Fold Cross Validation was used to tune the parameters along with a grid of hyperparameter sets. Parameter tuning process was done repeatedly with different subsets of features. Here the subsets were selected according to the forward feature selection method.

Finally the best feature set was selected by calculating the feature importance matrix. The most important features have a more impact on user behaviour.

3.4 Customer Trends Detection

User check-ins for a selected business can be viewed as a time series when it is plotted over time. The plotted graph can be divided into phases where each phase can be interpreted by a linear polynomial. In the context of time series these phases are called segments and each segment represents a trend in user check-ins. When the graph has a negative slope it symbolizes the pattern of customers leaving the business. On the other hand when graph has a positive slope it indicates that there is a trend of customers visiting that business. Time series segmentation techniques can be used to find trends and patterns in user check-ins graph.

Further we can analyse the review text posted during that period of trend and retrieve the trending topics which are popular among the crowd. This will reveal the reasons behind each trend whether it is because of the business or other external causes like seasonal trends. Also, more importantly what are the things that customers are interested in the particular business and characteristics that drive away the customers.

3.4.1 Finding Patterns Using Change Points

In order to find the patterns in user check-ins first user segments should be identified when user check-ins are plotted against time. Segmentation was done by analyzing the change points and the approach suggested by Taylor and Wayne [38]. The suggested is a combination of cumulative sum charts (CUSUM) and bootstrapping analysing.

First CUSUM chart was generated by the following steps.

1. If check-ins over time can be denoted as x1, x2, x3 … xn, Calculate the mean $\underline{x}$.

$$\bar{x} = \frac{x1+x2+\ldots+Xn}{n}$$

2. If the cumulative sum for the above data set can be denoted by S0, S1, S2 ...Sn, Calculate the cumulative Sum Si. Let S0=0

$$S_i = S_{i-1} + (xi - \bar{x}),\ i = 1, 2, 3, \ldots . n$$

3. Create a boundary for the CUSUM by estimating the magnitude of the changes.

$$S^i_{diff} = \max_{i=0,\cdots,n} S_i - \min_{i=0,\cdots,n} S_i = S_{max} - S_{min}$$

To analyze and recognize the changes in pattern using CUSUM charts need expertise and it should be performed manually. So Bootstrapping analysis has been introduced to mitigate this issue to identify change points systematically.

Bootstrap analysis was executed 1000 times repeatedly. Following are steps for a single

execution.

1. Shuffle the original data set and generate a new randomly ordered set. This is known as sampling without replacement (SWOR).
2. Calculate bootstrap CUSUM chart.
3. Calculate the magnitude of change for the bootstrap CUSUM which is denoted by Sjdiff.

After executing bootstrap analysis 1000 times, the following steps were carried out to determine whether there is a change in the data set.

1. Calculate the number of bootstraps where Sjdiff < SiDiff which is denoted by K.
2. Calculate the confidence level. Let N be the total number of bootstraps which is 1000.

$$\text{Confidence Level} = (K / N) \times 100$$

3. If confidence level is greater than 90% there is a change point in the data set.

To find out when the change has happened, The furthest point from the zero of the CUSUM chart can be calculated as follows.

$$|S_m| = \max_{i=0,\ldots,24} |S_i|$$

Once a change has been detected, the data set was splitted into two segments around the change point and the analysis repeated for each segment. Multiple change points were retrieved using this method.

Patterns were generated from the change points derived from the method in sections 3.4.1. A Pattern is a segment between two change points.

3.4.2 Extracting Trending Topics

To find out what are the most popular topics within a period, a frequent itemset mining approach was carried out. First reviews posted during each trend were retrieved. Then reviews were tokenized and each token represented an item in itemset while review being the itemset. The tokens underwent a preprocessing phase. Following are the text

preprocessing steps.

1. Turn in to lowercase
2. Removing punctuation marks
3. Removed white spaces
4. Removing stop words
5. Lemmatization

The algorithm that was used to derive trending topics is Apriori. The minimum support was varied between 0.5 and 0.7. Higher minimum support was used for a higher number of reviews. Lower number of reviews (<=2) were tested with minimum support of 0.5.

1-itemset, 2-itemset and 3-itemsets were generated using each algorithm. To enhance the insights derived from the output results, reviews were categorized and analyzed. Bigrams and trigrams also tried in order to find the most popular phrases during a time period. Experiment was conducted on monthly trending words.

4. RESULTS

4.1 Impact of Reviews on Customer Behaviour

4.1.1 Selecting Best Model

This section contains the model evaluation metrics for the classification. Table 4.1.1 contains the accuracy values for each model. Ensembled model was used with models, Random Forest and Xgboost. Precision, Recall and F1-score values are presented in Table 4.1.2 for the three models.

Table 4.1.1

Accuracy values of classification models

Model	Accuracy(%)
Logistic Regression	49.52
Neural Network	66.88
Random Forest (RF)	76.63
Xgboost (XGB)	81.31
Ensemble Classifier (RF, XGB)	81.14

Table 4.1.2

Precision, Recall and F1-score values of best performing classification models

Random Forest (RF)			
	Decrement (-1)	Neutral (0)	Increment (+1)
Precision	71.32	79.97	79.34
Recall	77.74	73.52	78.16
F1-score	74.39	76.61	78.75
Xgboost (XGB)			
	Decrement (-1)	Neutral (0)	Increment (+1)
Precision	79.23	83.46	81.45

Recall	79.39	81.47	82.88
F1-score	79.31	82.46	82.16
Ensemble Classifier (RF, XGB)			
	Decrement (-1)	Neutral (0)	Increment (+1)
Precision	78.83	83.13	81.65
Recall	79.81	80.11	83.16
F1-score	79.32	81.59	82.40

According to the table 4.1.1 Xgboost was performing well for the data set. However all the models were selected for parameter tuning to improve their results.

4.1.2 Hyper Parameter tuning

This section contains the model evaluation metrics after hyper parameter tuning for the best performing models. According to Table 4.1.3 most of the models have improved their performance except Logistic Regression. Even Though Neural Network and Random Forest have improved their accuracy a lot, Xgboost shows the best results.

Table 4.1.3

Accuracy values of best performing models after parameter optimization

Model	Accuracy(%)
Logistic Regression (LR)	49.61
Neural Network (NN)	75.09
Random Forest (RF)	82.47
Xgboost (XGB)	83.02
Ensemble Classifier (RF, XGB)	83.25

Table 4.1.4

Precision, Recall and F1-score values of best performing classification models after parameter optimization

Random Forest (RF)			
	Decrement (-1)	Neutral (0)	Increment (+1)
Precision (%)	74.09	71.77	80.62
Recall (%)	71.5	66.85	85.46
F1-score (%)	72.77	69.23	82.97
Xgboost (XGB)			
	Decrement (-1)	Neutral (0)	Increment (+1)
Precision (%)	72.63	72.53	83.59
Recall (%)	73	69.42	85.30
F1-score (%)	72.81	70.94	84.44
Ensemble Classifier (RF, XGB)			
	Decrement (-1)	Neutral (0)	Increment (+1)
Precision (%)	73.44	73.63	83.53
Recall (%)	74	69.42	85.78
F1-score (%)	73.72	71.47	84.64

4.1.3 Identifying Important Features

The performance of the best performing model was improved by forward selection method. A subset of features were selected as it can beat the accuracy of the current model with all the features. A subset of 18 features gave the maximum accuracy of 84%. The drop out features were 'neutral_review_count' and 'popularity'.

Table 4.1.5

Final results of the best performing model

Xgboost (XGB) **Accuracy : 84.05 %**			
	Decrement (-1)	Neutral (0)	Increment (+1)
Precision (%)	81.16	86.44	84.78

Recall (%)	82.50	82.61	86.58
F1-score (%)	81.82	84.48	85.67

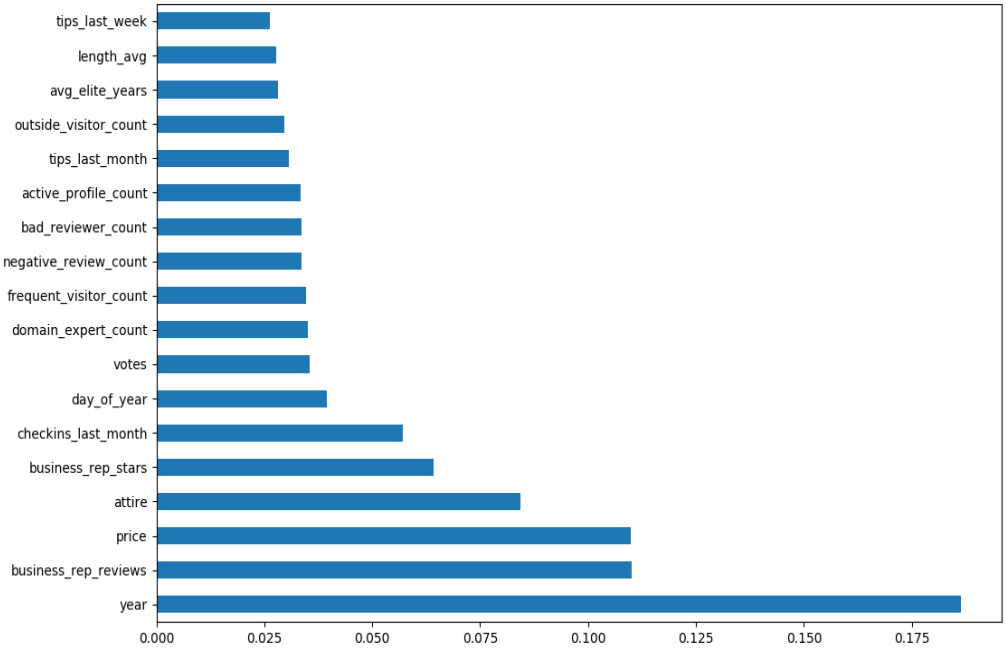
Table 4.1.6

Final results of best features set for the best performing model

Features of Reviewer	Features of the Review	Features of Business	Features of the Prediction date
avg_elite_years	length_avg	attire	year
frequent_visitor_count	votes	price	day_of_year
frequent_visitor_count		business_rep_reviews	
bad_reviewer_count		business_rep_stars	
outside_visitor_count		tips_last_week	
active_profile_count		tips_last_month	
domain_expert_count		checkins_last_month	
		negative_review_count	

Figure 4.1.1

Feature importance graph of best performing model Xgboost



4.2 Trends in Customer Behaviour

4.2.1 Long Term Trend Analysis

4.2.1.1 Change Point Detection

The most visited restaurant in Arizona, USA was selected and check-in throughout the year 2006 - 2018 was plotted against the number of days in each year. Linear Interpolation was used to get a rough idea about how check-ins have been changed over time. (Figure 4.2.1) Next, method described in section 3.4.1 was used to calculate change points. CUSUM chart and derived change points are shown in(Figure 4.2.2). Since bootstrap reordering does a random reordering, results for change points can be differ in each run. The time period between two consecutive change points was identified as a trend. By looking at the magnitude of the CUSUM chart the trends can be categorized into uptrend, downtrend or no change.

Figure 4.2.1

Daily check-ins in for most visited restaurant in Arizona, USA (2016-2018)

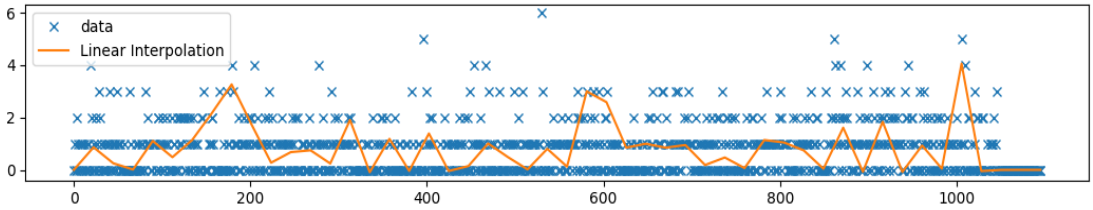


Figure 4.2.2

CUSUM chart and change points for most visited restaurant in Arizona, USA (2016-2018)

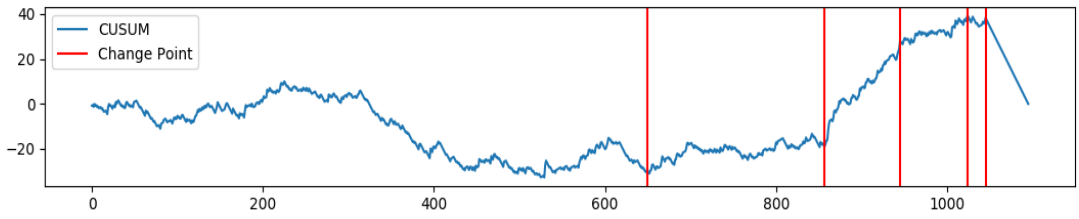
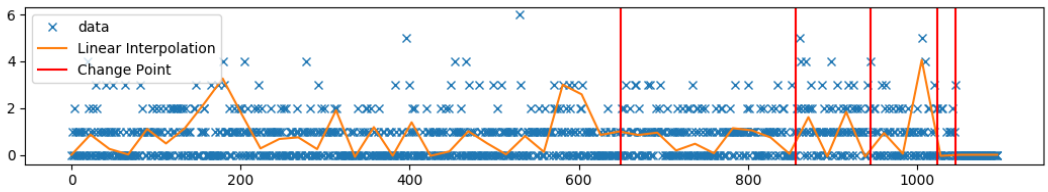


Figure 4.2.3

Daily check-ins and change points for most visited restaurant in Arizona, USA (2016-2018)



4.2.1.2 Trending Topic Extraction

According to the graph in figure 4.2.2, there are 5 periods that can be identified from the change points. Frequent Itemset mining approach was applied to the reviews posted in each trending period. Table 4.2.1 shows the frequent items sets found in each period. It shows that the menu item 'taco' is famous since 2016 and it has maintained its popularity all the time. Since this is appeared in all the time period it can be identified as a restaurant speciality or all time favourite and trending menu item.

Table 4.2.1

Trending topics of most visited restaurant in Arizona, USA (2016-2018)

Time Period	Number of Reviews	Overall Sentiment of Reviews	Apriori (1 itemset)
2016-01-01 2017-10-12	467	Positive	taco
2017-10-12 2018-05-07	171	Positive	taco
2018-05-07 2018-08-03	110	Positive	taco
2018-08-03 2018-10-21	74	Positive	Place, taco
2018-10-21 2018-11-11	14	Positive	Place, taco
2018-11-11 2018-12-31	3	Positive	Food, love

4.2.2 Short Term Trend Analysis

4.2.2.1 Change Point Detection

The most visited restaurant in Arizona, USA was selected and checked-in throughout the year 2018 to analyse the short term trends. CUSUM chart and derived change points are shown in Figure 4.2.4. CUSUM chart and derived change points for month of May

in 2018 is shown in Figure 4.2.5

Figure 4.2.4

CUSUM chart and change points for most visited restaurant in Arizona, USA (2018)

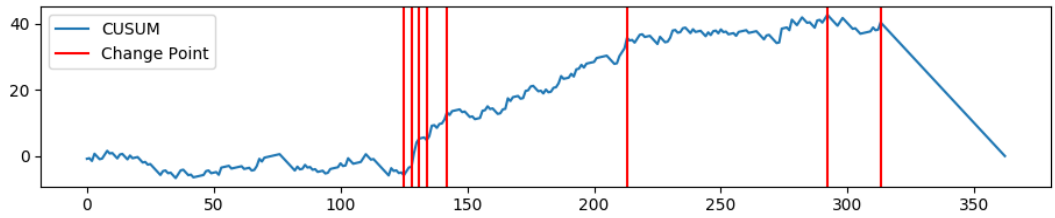
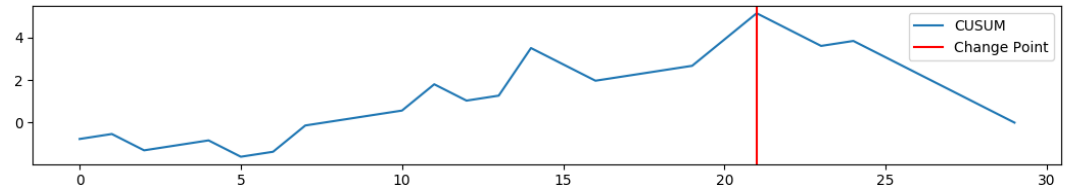


Figure 4.2.5

CUSUM chart and change points for most visited restaurant in Arizona, USA (April 2018)



4.2.2.2 Trending Topic Extraction

According to the graph in figure 4.2.4, there are 9 segments that can be identified from the change points. Table 4.2.2 shows the frequent items sets found in each period. There are more trending periods in the chart and there are different trending words in each period. According to the graph in figure 4.2.5 there is a change point in April which breaks the ongoing trend and there are interesting terms found in each period.

Table 4.2.2

Trending topics of most visited restaurant in Arizona, USA

(2016-2018)

Time Period	Number of Reviews	Overall Sentiment of Reviews	Apriori (1 itemset)
2018-01-02 2018-05-07	98	Positive	taco
2018-05-07 2018-05-10	4	Positive	Good, great, ive, taco, try, wait
2018-05-10 2018-05-13	10	Positive	Food, good, place, taco
2018-05-13 2018-05-16	4	Positive	Backyard, food, great, place, taco
2018-05-16 2018-05-24	12	Positive	Food, taco
2018-05-24 2018-08-03	80	Positive	taco
2018-08-03 2018-10-21	74	Positive	Place, taco
2018-10-21 2018-11-11	14	Positive	Place, taco
2018-11-11 2018-12-30	3	Positive	Food, love

Table 4.2.3

Trending topics of most visited restaurant in Arizona, USA

(April 2018)

Time Period	Number of Reviews	Overall Sentiment of Reviews	Apriori (1 itemset)
2018-04-01 2018-04-22	20	Positive	Food, taco
2018-04-22 2018-04-30	3	Positive	Fresh, love, make, meal, plenty, salsa, taco, time

4.2.3 Periodic Trending Topics

Trending topics can be derived periodically other than the trends. As shown in sections 4.2.1 and 4.2.2 if the period is short more informative topics can be examined. Table 4.2.4 shows the monthly trending topics for four selected restaurants.

Table 4.2.4

Monthly Trending Trending topics of Four Selected restaurant in Arizona, USA (2018)

Month in 2018	Business 1 (Number of reviews)	Business 2 (Number of reviews)	Business 3 (Number of reviews)	Business 4 (Number of reviews)
January	Broth, good, hachi, place, ramen, service, shoyu (14)	dinner, dish, eve, food, great, new, prompt, really, service, year, yelp (4)	-	food, good, meatball (93)
February	Broth, great, owner, place, ramen (10)	Service, time (7)	-	food, meatball (78)
March	Ramen (14)	Delicious, food (12)	-	food, great, meatball (84)
April	Food, place, ramen (7)	Good, like, place, service (6)	food, new,wildflower (11)	food, meatball (96)
May	Place, ramen (17)	Delicious, food (7)	food, good, new (8)	food, meatball (79)
June	Broth, delicious, ramen (18)	definitely, dinner, food, good, great, night, old, recently, restaurant, service, sure, town, visit (3)	location (5)	food, meatball (87)
July	Ramen (17)	amazing, food, location, place, scottsdale, service, try (4)	eat (2)	food, meatball (76)
August	Best, great, ramen (19)	al, best, dente, didnt, dinner, fish, food, friendly, gem, good, hidden,husband, meal, perfect, place, restaurant, say, scottsdale, staff, truly (4)	day, food, great, sandwich,service , time, wildflower, sandwich (8)	food, meatball (72)
September	Ramen (19)	Delicious, food, great, italian, lobster, old, risotto (7)	good, linda, nice, order,really, wildflower (6)	42 food, good, great, meatball, place, service (42)
October	Ramen (15)	best, calamari, food, italian, little, restaurant, risotto, sauce, service (6)	better,delicious,good,item,option,tea,way (3)	food, great, meatball (64)

November	Food, really (2)	Delicious, just (5)	-	food, great, meatball (28)
December	(0)	(0)	-	-

Generating 2-itemset and 3-itemsets were not much effective as they contained the same items from 1 itemset. Most of the time empty lists were the outputs. Bigrams and trigrams were generated without removing the stop words in order to preserve the meaning. Only the common phrases were able to be extracted when performing frequent itemset mining. No meaningful phrases found. When bigrams and trigrams were generated after removing stop words some meaningful phases could be found. But not always. Extracting frequent words for good and bad reviews separately was more effective. Decisions can be taken whether the topics are popular for good or bad.

Table 4.2.4

Frequent Itemsets derived by N-grams

Month in 2018	Business 2 Unigram	Business 2 Bigram	Business 2 Trigram
January	dinner, dish, eve, food, great, new, prompt, really, service, year, yelp	(new, year), (year, eve)	(new, year, eve)
June	definitely, dinner, food, good, great, night, old, recently, restaurant, service, sure, town, visit	(old, town)	-
August	al, best, dente, didnt, dinner, fish, food, friendly, gem, good, hidden,husband, meal, perfect, place, restaurant, say, scottsdale, staff, truly	(al, dente), (hidden, gem), (staff, friendly)	-

Table 4.2.5

Frequent Itemsets derived for review categories

Month in 2018	Business 2 Good Reviews (Number of reviews)	Business 2 Bad Reviews (Number of reviews)
January	Great (2)	Eve, food, new, year,yelp (2)
April	Like, place, service (5)	“Food was not good, the server was rude, and they charge for ice in your drinks!” (1)
May	Delicious, food, sauce, service (6)	“ We ordered chicken Parmesan and gnocchi. The homemade gnocchi was horrible and the chicken Parmesan was just as bad” (1)
August	Best, friendly, gem, good, hidden, husband, meal, restaurant, say, scottsdale, staff, truly, went (3)	“We started with the grilled Calamari, was burnt and needed a steak knife to cut it, once you got past the burnt taste it was worse than chewing rubber” (1)

5. DISCUSSION

5.1 Impact on User Behaviour

According to the results in section 4.1 it is evident that user behaviour can be predicted from the reviews posted by the yelp users. According to the feature importance matrix, features describing the operational level and the business category. Following are these features according to their importance,

1. Total reviews of the business
2. Price
3. Attire
4. Star rating of the business
5. Check-ins last month

Review date which is denoted by year and day of the year, also plays a major role when predicting user behaviour. This is probably because, Factors like seasonal climate changes and festive seasons that influence the customer can be derived from the review date. Also Popularity or the operation level of the business is changing overtime.

Features of the reviewer and review comes after that. The most prominent features of a reviewer that can predict users visiting a business are following according to their importance.

1. Domain expert
2. Frequent visitor
3. Reviewer type (Bad)
4. Activeness
5. Outside visitor
6. Number of elite years

Being a domain expert and frequent visitor is the most important feature of a reviewer. In this study, domain experts are the reviewers who write about the same kind of businesses. The reviewer can compare and give a fair comment on a business when the review is visiting businesses belonging to a common category. Frequent visitor is someone who has many reviews about the same business. This means the reviewer has

visited the business multiple times.

If a reviewer is giving bad comments often he can be recognized as a bad reviewer. According to the dataset most of the reviewers are good. They post good reviews most of the time. But there are few who post a review about the business only if they get a bad experience. In this study active users are reviewers who are posting reviews often on businesses in recent months . Usually most of the businesses reviewed by a reviewer are located where he lives. When a business gets a review from an outsider who is not a local resident around the business area the model finds it important. The last feature is the number of years that the reviewer is entitled as ‘elite’. Yelp gives the title to the users who are active through the year.

Following are some removed features of the reviewer which are not influential on user behaviour.

1. Popularity - Number of fans and friends that the reviewer has
2. Being Elite on the prediction date
3. Legacy profile - Having a long history as a Yelp user

Features of review that are important are votes received by the review, length of the review and sentiment of the review. According to the dataset, the number of negative reviews received by a company is very few compared to positive reviews. So getting a negative review affects user behaviour.

Some important features that were identified are the number of tips received by the business last week and last month. Tips are more shorter comments than the review. But they provide useful information. Because of the length of the tips they are easy to read and catch the idea. It is interesting features regarding the number of tips are there in the selected feature list over the features regarding the number of reviews. This might be an indication of users preference tips.

5.2 Trends and Trending Topics

Capturing trends in customer visits is successful according to the graphs. Proposed method is very precise and accurate when detecting change points. It is much more robust to outliers than spline linear interpolation. This approach performs well with both

long term and short term trends.

Long term trending topic analysis is more suitable to derive more generalized ideas about the business. Short term trending topic analysis is more appropriate to derive more informative and detailed topics. According to the results the most visited restaurant in Arizona, USA from 2016 to 2018, has only 5 time segments during 3 years. The generated topics are not unique to the individual time period. But it gives a hint about what is the business's most popular for all the time. Overall image of a business can be gotten from long trending topic analysis.

When considering only 1 year of data for above business, there were 9 segments visible. Derived topics are more descriptive than long term trending topics. This might be because of the number of reviews that the trending period has. Long time periods have many reviews and in order to derive meaningful itemsets minimum support had to decrease. When there is a small number of reviews, minimum support is increased. The time period should not be too short; there should be enough data; user checkins to plot and user reviews to derive topics.

Evaluating trending topics periodically is very beneficial to the business. Specially the popular menu items can be easily identified with this approach. According to the results of long term and short term trend analysis, short periods could output more informative itemsets. So monthly trending topics were derived and analyzed for few businesses. Different itemsets were output for each month.

When interpreting these results knowing about the actual context of the business is important. Otherwise there can be misleading information. For that, going through the actual comments was done and verified the interpretations. As an example long term trending topics have the term 'taco'. When going through the actual reviews it was found that most of the reviews compared to taco from the selected business with 'Taco Bell'. Taco Bell is a famous food chain which specializes in tacos. And when considering the 5th trend in short term trend analysis, the term 'backyard' is there. For instance we think that this might be an attribute of the business. But when going through the actual comments it was found that the term is in restaurants name and reviewers are referring to the restaurant by it's name.

When analysing periodic trending topics in January, business 1 has many reviews talking about New year's evening dinner when going through the reviews. Hopefully the method has itemsets containing words New, year, eve and dinner. But when analysing review categories separately these keywords fell into Bad reviews. Which means most customers had a bad experience at the end of december last year. Following are some reviews.

"Came with a large group of friends prior to New Years Eve"

"We went there for New Years Eve dinner and the food was not worth the price"

Same with September and October. Restaurants have higher demand for italian food like risotto, calamari, lobster and sauce. And all the reviews are good. Following are some reviews.

"The bolognese was unbelievable and my seafood risotto was the best I've ever had"

"We ordered the calamari in Limoncello sauce and dipped the bread in it"

There are some common words that can be identified in reviews. They are Service, food, place, location & etc. This indicates that these are the general topics that the reviewers are talking about the reatuarants. Which means reviews are not describing the other external factors like weather, factors that affect on the customer visits. So finding about the real causes behind the trends are difficult. If we can integrate these reviews with social media posts and news articles, the interpretation can be improved. But trending menu items in a given period can be retrieved successfully from this approach.

6. CONCLUSION

According to the results of this study it is evident that user reviews have an impact on user behaviour. When predicting the user behaviour whether the user visits are going to increase, decrease or have no change with respect to last month stats, features that are describing the operational level and the business category are the most important factors. Features of the reviewer and review were studied along with these features. The best performing classification algorithm that fitted the dataset was Xgboost with accuracy of 84%.

Domain expersity and activeness are the most important features of the reviewer. A bad reviewer's review has more impact than a good reviewer. Frequent visitor’s reviews can be more influential than first time reviewer’s posts. An outsider to the business location is more important than a local resident. Being entitled as ‘elite’ on the prediction date, having many fans, having many friends and having a profile with a long history is not very important. But the number of years that the reviewer got ‘elite’ matters. This reflects the reviewer’s long time contribution towards Yelp is more important than reviewer reputation and popularity. Moreover, Tips play an important role. This is most probably because of its length. Users might prefer more short useful reviews. Votes received by a review is also an important feature. This is an indicator of users' response towards the review.

In the future different subset of features can be tried with different algorithms. Also this approach can be applied on different data sets in different domains. Since the features that are introduced in this study are not specific to restaurants, predicting customer behaviour in other business domains using suggested features is highly encouraged.

When analysing the trends, the applied method with CUSUM chart and Bootstrap sampling is precise when finding change points. Frequent Itemset set mining also works well with the text when deriving common topics in reviews. Long term trend analysis and long term trending topics gave a generalized view of the business. Short term trend analysis and short term trending topics were more descriptive about each trending period. By using that fact, trending topics that were generated periodically for short time periods. The results are very useful inorder to find out the common topic among

customers. If there is sufficient data, the short term trend analysis can be performed periodically on a monthly basis. This is an advantive way for a business to evaluate it's services and popularity.

Currently meaningful phrases can be extracted by performing frequent itemset mining on bigrams and trigrams for good and bad reviews separately. But when bigrams and trigrams do not return a frequent pattern, interpreting the results should be done by someone who is familiar with the business. In future this study can be extended to output meaningful phrases that can be constructed by trending topics while maintaining the common idea behind the reviews. So interpretation can be done automatically.

There are a set of common words identified in the study. They are Service, Food, Place and Location. Since the reviewers are only talking about these areas in their reviews, retrieved topics can’t be used to find out the reasons behind the customer trends alone. In the future review trending topics should be integrated with other external trending topics from news and social media posts in order to identify the real causes behind the trend. Also these words can be removed in text preprocessing steps to retrieve more unique topics. These words will be useful in entity tagging in reviews of restaurants.

REFERENCES

1. *Lian, Jianxun, et al. "Restaurant survival analysis with heterogeneous information." Proceedings of the 26th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee, 2017.*
2. *Luca, Michael. "Reviews, reputation, and revenue: The case of Yelp. com." Com (March 15, 2016). Harvard Business School NOM Unit Working Paper 12-016 (2016).*
3. *Luca, Michael, and Georgios Zervas. "Fake it till you make it: Reputation, competition, and Yelp review fraud." Management Science 62.12 (2016): 3412-3427.*
4. *Lim, Young-shin, and Brandon Van Der Heide. "Evaluating the wisdom of strangers: The perceived credibility of online consumer reviews on Yelp." Journal of Computer-Mediated Communication 20.1 (2014): 67-82.*
5. *Altenburger, Kristen M., and Daniel E. Ho. "Is Yelp Actually Cleaning Up the Restaurant Industry? A Re-Analysis on the Relative Usefulness of Consumer Reviews." The World Wide Web Conference. ACM, 2019.*
6. *Gunden, Nefike. "How Online Reviews Influence Consumer Restaurant Selection." (2017).*
7. *Lei, Mingtao, Lingyang Chu, and Zhefeng Wang. "Mining top-k sequential patterns in database graphs: a new challenging problem and a sampling-based approach." arXiv preprint arXiv:1805.03320 (2018).*
8. *Hu, Minqing, and Bing Liu. "Mining and summarizing customer reviews." Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004.*
9. Sun, Yanmin, Andrew KC Wong, and Mohamed S. Kamel. "Classification of imbalanced data: A review." *International Journal of Pattern Recognition and Artificial Intelligence* 23.04 (2009): 687-719.
10. Kuncheva, Ludmila I., et al. "Instance selection improves geometric mean accuracy: a study on imbalanced data classification." *Progress in Artificial Intelligence* 8.2 (2019): 215-228.
11. Wang, Shoujin, et al. "Training deep neural networks on imbalanced data sets." *2016 international joint conference on neural networks (IJCNN)*. IEEE, 2016.
12. Lin, Enlu, Qiong Chen, and Xiaoming Qi. "Deep Reinforcement Learning for

Imbalanced Classification." *arXiv preprint arXiv:1901.01379* (2019).

13. Aminikhanghahi, Samaneh, and Diane J. Cook. "A survey of methods for time series change point detection." *Knowledge and information systems* 51.2 (2017): 339-367.
14. Kawahara, Yoshinobu, and Masashi Sugiyama. "Sequential change-point detection based on direct density-ratio estimation." *Statistical Analysis and Data Mining: The ASA Data Science Journal* 5.2 (2012): 114-127.
15. Esling, Philippe, and Carlos Agon. "Time-series data mining." *ACM Computing Surveys (CSUR)* 45.1 (2012): 12.
16. Keogh, Eamonn, and Shruti Kasetty. "On the need for time series data mining benchmarks: a survey and empirical demonstration." Data Mining and knowledge discovery 7.4 (2003): 349-371.
17. Keogh, Eamonn, et al. "Segmenting time series: A survey and novel approach." *Data mining in time series databases*. 2004. 1-21.
18. Keogh, Eamonn, et al. "An online algorithm for segmenting time series." *Proceedings 2001 IEEE International Conference on Data Mining*. IEEE, 2001.
19. Chung, Fu-Lai, et al. "An evolutionary approach to pattern-based time series segmentation." *IEEE transactions on evolutionary computation* 8.5 (2004): 471-489.
20. Yu, Jingwen, et al. "A pattern distance-based evolutionary approach to time series segmentation." *Intelligent Control and Automation*. Springer, Berlin, Heidelberg, 2006. 797-802.
21. Fu, Tak-chung, et al. "A specialized binary tree for financial time series representation." 3 rd International Workshop on Mining Temporal and Sequential Data (TDM-04). 2004.
22. Fu, Tak-chung, Fu-lai Chung, and Chak-man Ng. "Financial Time Series Segmentation based on Specialized Binary Tree Representation." *DMIN* 2006 (2006): 26-29.
23. *Zhang, Kunpeng, Ramanathan Narayanan, and Alok N. Choudhary. "Voice of the Customers: Mining Online Customer Reviews for Product Feature-based Ranking." WOSN 10 (2010): 11-11.*
24. *Hu, Minqing, and Bing Liu. "Mining and summarizing customer reviews." Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004.*
25. *Xu, Shishuo, et al. "TRAFFIC EVENT DETECTION USING TWITTER DATA*

BASED ON ASSOCIATION RULES." ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences 4 (2019).

26. *Maylawati, D. S. "The concept of frequent itemset mining for text." IOP Conference Series: Materials Science and Engineering. Vol. 434. No. 1. IOP Publishing, 2018.*
27. Anenberg, Elliot, Chun Kuang, and Edward Kung. *Social Learning and Local Consumption Amenities: Evidence from Yelp. Working paper, 2018*
28. *Farhan, Wael. "Predicting Yelp Restaurant Reviews." UC San Diego, La Jolla (2014)*
29. *Bakhshi, Saeideh, Partha Kanuparthy, and Eric Gilbert. "Demographics, weather and online reviews: A study of restaurant recommendations." Proceedings of the 23rd international conference on World wide web. ACM, 2014*
30. *Nakayama, Makoto, and Yun Wan. "The cultural impact on social commerce: A sentiment analysis on Yelp ethnic restaurant reviews." Information & Management 56.2 (2019): 271-279.*
31. *Carbon, Kyle, Kacyn Fujii, and Prasanth Veerina. "Applications of machine learning to predict Yelp ratings." (2014).*
32. *Fan, Mingming, and Maryam Khademi. "Predicting a business star in yelp from its reviews text alone." arXiv preprint arXiv:1401.0864 (2014).*
33. *Asghar, Nabiha. "Yelp dataset challenge: Review rating prediction." arXiv preprint arXiv:1605.05362 (2016).*
34. *Kong, Angela, Vivian Nguyen, and Catherina Xu. "Predicting international restaurant success with yelp." (2016).*
35. *Shen, Ruhui, et al. "Predicting usefulness of Yelp reviews with localized linear regression models." 2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS). IEEE, 2016.*
36. *Liu, Xinyue, Michel Schoemaker, and Nan Zhang. "Predicting Usefulness of Yelp Reviews." (2016).*
37. *Arif, Siti Nur Afiqah Mohd, et al. "Change point analysis: a statistical approach to detect potential abrupt change." Jurnal Teknologi 79.5 (2017).*
38. *Taylor, Wayne A. "Change-point analysis: a powerful new tool for detecting changes." (2000): 01-01.Kass-Hout, Taha A., et al. "Application of change point analysis to daily influenza-like illness emergency department visits." Journal of the American Medical Informatics Association 19.6 (2012): 1075-1081.*

APPENDICES

[Appendix - I : New features derived from original data set for review]

Old Features *Entity : feature (data type)*	**New Feature**	**Description**
Review : stars (decimal)	-	Star rating for the review
Review : stars (decimal)	Review_type (string)	Whether it's a good review, bad review or neutral review based on review stars If stars >= 4 then type is good If stars <=2 then type is bad Otherwise type is neutral
Review : useful (int)	-	Useful votes received by the review
Review : funny (int)	-	Funny votes received by the review
Review : cool (int)	-	Cool votes received by the review
Review : Text (string)	Review_length (int)	Number of words in the text
Review : Text (string)	Sentiment_score (string)	Score to indicate the sentiment of the text If score > 0 then sentiment is positive If score < 0 then sentiment is negative (library - nltk.sentiment.vader)
Review : Text (string) Business : Attributes (array of attributes)	*Business_attributes_similarity (decimal)	Jaccard similarity score between business attributes and the review text
Review : Text (string) Business : Categories (array of categories)	*Business_domain_similarity (decimal)	Jaccard similarity score between business categories and the review text

* - New features that will be introduced in this research

[Appendix - II : New features derived from original data set for reviewer]

Old Features *Entity : feature (data type)*	New Feature	Description

Reviewer : friends (array of user_ids)	friends (int)	Number of friends that the reviewer has
Reviewer : fans (int)	-	Number of fans
Reviewer : useful, funny, cool	*votes_sent (int)	Number of votes sent by the reviewer
Reviewer : compliment_hot, compliment_more, compliment_profile, compliment_cute, compliment_list, compliment_plain, compliment_cool, compliment_funny, compliment_writer, compliment_photos (int)	*compliment_received (int)	Number of compliments received by the reviewer
Reviewer : yelping_since (date)	*yelp_age (int)	Number of days from yelping_since up to prediction date
Reviewer : elite (array of years)	is_elite (boolean)	Is reviewer is marked as ‘elite’ in the year of the prediction date
Reviewer : elite (array of years)	elite_years (int)	Number of years the user is entitled as ‘elite’
Reviewer : yelping_since (date)	*legacy_profile (boolean)	User profiles which has a long history If yelp_age >= 5 years
Reviewer Reviews : stars (decimal)	*reviewer_type (string)	Whether the reviewer is a good, bad or neutral reviewer based on average stars given to business from the beginning up to prediction date If avg stars >= 4 then type is good If avg stars <=2 then type is bad Otherwise type is neutral
Reviewer Reviews : date (date)	total_review_count (int)	Review count of the reviewer up to prediction date
Reviewer Reviews : date (date)	*active_profile (boolean)	Reviewer is posting reviews regularly throughout the time (If reviewer has posted at least one review per month in past three months)

Reviewer Reviews : date (date)	*Recently_active_profile (boolean)	Reviewer is posting many reviews recently (If reviewer has posted more than one review within past month)
Reviewer Reviews : date (date) Business	*first_time_reviewer (boolean)	Reviewer reviews about a particular business for the first time
Reviewer Reviews : date (date) Business	*regular_reviewer (boolean)	Reviewer had reviewed about the same business multiple times
Reviewer Reviews : date (date) Business : city (string)	*local_reviewer (boolean)	Reviewer is from business area If $\frac{Reviews\ postested \in same\ business\ city}{Total\ review\ count} >= 0.5$
Reviewer Reviews : date (date) Business : state (string)	*outside_reviewer (boolean)	Reviewer is from other area If $\frac{Reviews\ postested \in same\ business\ state}{Total\ review\ count} < 0.5$
Reviewer Reviews : date (date) Business : categories (string)	*domain_expert (boolean)	Reviewer has many reviews on same kind of business (High jaccard similarity between business category and the all the reviews posted up to date)

* - New features that will be introduced in this research

[Appendix - III : New features derived from original data set regarding business status]

Old Features *Entity : feature (data type)*	New Feature	Description
Business Reviews	business_rep_reviews	Total reviews that the business has
Business : stars (decimal)	business_rep_stars	Business star ratings
Business: attributes.attire (string)	attire	Casual, Dressy, Formal
Business: attributes.price_range (int)	price	Price Range 1 to 5

Business Review : date (date)	*reviews_last_month (int)	Reviews received by business during last month
Business Check-ins : date (date)	*checkins_last_month (int)	Check-ins received by business during last month
Business Tips : date (date)	*tips_last_month (int)	Tips received by business during last month
Business Review : date (date)	*reviews_last_week (int)	Reviews received by business during last week
Business Check-ins : date (date)	*checkins_last_week (int)	Check-ins received by business during last week
Business Tips : date (date)	*tips_last_week (int)	Tips received by business during last week

* - New features that will be introduced in this research

[Appendix - IV : Features of the training set based on features of the reviewer]

Entity: Feature	**Feature**	**Description**
Filtration : All the reviewers that has posted reviews for a business during last month before the prediction date		
Reviewer : fans	avg_fans	Average number of fans that a reviewer has
Reviewer : elite_years	avg_elite_years	Average number of years that a reviewer was entitled as ‘elite’
Reviewer : is_elite	elite_profile	Number of elite profile count
Reviewer : regular_reviewer	regular_reviewer_count	Number of regular reviewers
Reviewer : first_time_reviewer	first_time_reviewer_count	Number of first time reviewers
Reviewer : reviewer_type	good_reviewer_count	Number of good reviewers If reviewer_type is good
Reviewer : reviewer_type	bad_reviewer_count	Number of bad reviewers If reviewer_type is bad
Reviewer : reviewer_type	neutral_reviewer_count	Number of neutral reviewers If reviewer_type is neutral
Reviewer :	local_reviewer_count	Number of local reviewers

local_reviewer		
Reviewer : outside_reviewer	outside_reviewer_count	Number of outside reviewers
Reviewer : legacy_profile	legacy_profile_count	Number of legacy profiles
Reviewer : active_profile	active_profile_count	Number of active profiles
Reviewer : recently_active_profile	recently_active_profile_count	Number of recently active profiles
Reviewer : domain_expert	domain_expert_count	Number of domain experts

[Appendix - V : Features of the training set based on features of the reviews]

Entity: Feature	**Feature**	**Description**
Filtration : All the reviews that has posted for a business during last month before the prediction date		
Review : Business_attributes_similarity	business_attribute_reviews	Number of reviews talking about business attributes
Review : stars	business_stars_avg	Average number star ratings given to the business
Review : review_type	good_review_count	Number of good reviews If review_type is good
Review : review_type	bad_review_count	Number of bad reviews If review_type is bad
Review : review_type	neutral_review_count	Number of neutral reviews If review_type is neutral
Review : funny	funny_vote_count	Number of funny votes received
Review : cool	cool_vote_count	Number of cool votes received
Review : useful	useful_vote_count	Number of useful votes received
Review : sentiment_score	positive_review_count	Number of positive reviews
Review : sentiment_score	negative_review_count	Number of negative reviews

Tips : sentiment_score	positive_tip_count	Number of positive tips
Tips : sentiment_score	negative_tip_count	Number of negative tips