

**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 Chithranjan

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

DECLARATION	i
ACKNOWLEDGEMENTS	ii
Abstract	iii
TABLE OF CONTENT	iv
LIST OF FIGURES	vii
LIST OF TABLES	viii
LIST OF ABBREVIATION	ix
LIST OF APPENDICES	x
INTRODUCTION	1
1.1 Yelp Data Set	2
1.1.1 Yelp.com	2
1.1.2 Yelp Challenge	3
1.1.3 Yelp Data	3
1.2 Significant Changes in User Behaviour	4
1.2.1 Identification of Changes	4
1.2.2 Predict future changes	4
1.3 Trends in Customer Behaviour	5
1.3.1 Customer Trends Detection	5
1.3.2 Reasons Behind Customer Trends	5
1.4 Problem Statement	5
1.5 Motivation	6
1.6 Objectives	6
1.7 Research Scope	7
LITERATURE REVIEW	8
2.1 Impact of customer reviews	8
2.1.1 Effect on Business Survival	8
2.1.2 Effect on Business Revenue	9
2.1.3 Effect on Customer Selection	10
2.1.4 Effect on Customer Ratings	10
2.1.5 Usefulness of Reviews	11
2.2 Perception of Customer Reviews	12
2.2.1 Familiarity with the platform	12
2.2.2 Learning rate of the users	12
2.2.3 Cultural influences	13

2.2.4 Customer trends	13
2.3 Features of Online Reviews	14
Table 2.3.1 Features of a Restaurant	14
Table 2.3.2 Features of a Review	15
2.4 Predicting User Behaviour using Classification	16
2.4.1 Classification	16
2.4.2 Imbalanced Data	17
2.4.3 Binary Class Classification on Imbalanced Data	18
2.4.3.1 Converting into Balanced Data	18
2.4.3.2 Algorithmic Approches	19
2.4.4 Multi Class Imbalanced Classification	20
2.5 Exploring Trends in Customer Behaviour	20
2.5.1 Time Series Data Mining	20
2.5.2 Time Series Segmentation	21
2.5.2.1 Identify Trends Using Time Series Segmentation	21
2.5.2.2 Identify Segments in Timeseris	21
2.5.2.3 Representation of Time Series Data	22
2.5.2.4 Piecewise Linear Approximation	22
2.5.3 Change Point Detection	23
2.5.4 Perceptually Important Point Detection	24
2.6 Extract Trending Topics in Reviews	26
2.6.1 Structured Version of Text	26
2.6.2 Frequent Itemset Mining vs Association Rule Mining	26
2.6.3 Support and Confidence	27
2.6.4 Frequent Itemset Mining in Text	27
2.6.4.1 Text Summarization Using Frequent Pattern Mining	28
METHODOLOGY	29
3.1 Experiment	29
3.2 Data	29
3.3 Understanding the Data Set	31
3.4 Predicting User Behaviour	38
3.4.1 Feature Selection	39
3.4.2 Class Labels	41
3.4.3 Annotating the Training Set	42
3.3.4 Training Data Set	43
3.3.5 Classification and Model Evaluation	44
3.4 Customer Trends Detection	45
3.4.2 Extracting Trending Topics	47
RESULTS	49
4.1 Impact of Reviews on Customer Behaviour	49
4.1.1 Selecting Best Model	49

4.1.2 Hyper Parameter tuning	50
4.1.3 Identifying Important Features	52
4.2 Trends in Customer Behaviour	53
4.2.1 Long Term Trend Analysis	53
4.2.1.1 Change Point Detection	53
4.2.1.2 Trending Topic Extraction	54
4.2.2 Short Term Trend Analysis	55
4.2.2.1 Change Point Detection	55
4.2.2.2 Trending Topic Extraction	56
4.2.3 Periodic Trending Topics	58
DISCUSSION	61
CONCLUSION	66
REFERENCES	68
APPENDICES	72
[Appendix - I : New features derived from original data set for review]	72
[Appendix - II : New features derived from original data set for reviewer]	73
[Appendix - III : New features derived from original data set regarding business status]	74
[Appendix - IV : Features of the training set based on features of the reviewer]	75
[Appendix - V : Features of the training set based on features of the reviews]	76

LIST OF FIGURES

Figure 2.5.4.1	Pseudo code of the PIP identification process	27
Figure 3.2.2	Relationships among Yelp Data set entities	33
Figure 3.3.1	Number of restaurants in each state in the USA	34
Figure 3.3.2	Number of Reviews, Check-ins, and Tips for selected five businesses in 2018	35
Figure 3.3.3	Monthly check-in count, review count and tips count over time for business “Fremont Street Experience”	36
Figure 3.3.4	Relationship between good review count before a certain date and customer check-ins after that date	37
Figure 3.3.5	Relationship between bad review count before a certain date and customer check-ins after that date	38
Figure 3.3.6	Monthly check-in count before and after a certain date over the time for a business	38
Figure 3.3.7	Difference of check-in counts before and after a certain date over the time for a business	39
Figure 3.3.8	Distribution of check-in difference	40
Figure 3.4.1	Correlation matrix between selected features	43
Figure 3.4.2	Distribution of the check-in difference	45
Figure 4.1.1	Feature importance graph of best performing model Xgboost	55
Figure 4.2.1	Daily check-ins in for most visited restaurant in Arizona, USA (2016-2018)	56
Figure 4.2.3	Daily check-ins and change points for most visited restaurant in Arizona, USA (2016-2018)	56
Figure 4.2.4	CUSUM chart and change points for most visited restaurant in Arizona, USA (2018)	58
Figure 4.2.5	CUSUM chart and change points for most visited restaurant in Arizona, USA(April 2018)	58

LIST OF TABLES

Table 2.3.1	Features of a Restaurant	15
Table 2.3.2	Features of a Review	16
Table 2.3.3	Features of a Reviewer	17
Table 3.2.1	Yelp data set	32
Table 3.3.1	Basis stats of the dataset	33
Table 3.4.1	Features of the training set	42
Table 4.1.1	Accuracy values of classification models	51
Table 4.1.2	Precision, Recall and F1-score values of best performing classification models	51
Table 4.1.3	Accuracy values of best performing models after parameter optimization	53
Table 4.1.4	Precision, Recall and F1-score values of best performing classification models after parameter optimization	53
Table 4.1.5	Final results of the best performing model	54
Table 4.1.6	Final results of best features set for the best performing model	54
Table 4.2.1	Trending topics of most visited restaurant in Arizona, USA (2016-2018)	57
Table 4.2.2	Trending topics of most visited restaurant in Arizona, USA (2016-2018)	59
Table 4.2.3	Trending topics of most visited restaurant in Arizona, USA (April 2018)	59
Table 4.2.4	Monthly Trending Trending topics of Four Selected restaurant in Arizona, USA (2018)	60
Table 4.2.4	Frequent Itemsets derived by N-grams	62
Table 4.2.5	Frequent Itemsets derived for review categories	62

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

Appendix I	New features derived from original data set for review	75
Appendix II	New features derived from original data set for reviewer	75
Appendix III	New features derived from original data set regarding business status	77
Appendix IV	Features of the training set based on features of the reviewer	78
Appendix V	Features of the training set based on features of the reviews	79