

Implicit Feature Extraction from Customer Reviews Using Supervised Learning

Atheesan Sornalingam

179306F

M.Sc. in Computer Science

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

March 2020

Implicit Feature Extraction from Customer Reviews Using Supervised Learning

Atheesan Sornalingam

179306F

This dissertation submitted in partial fulfillment of the requirements for the Degree of M.Sc. in Computer Science specializing in Data Science, Engineering, and Analytics

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

March 2020

DECLARATION

I declare that this is my own work and this PG Diploma Project Report does not incorporate without acknowledgment 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 acknowledgment 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.

.....

Atheesan Sornalingam

.....

Date

I certify that the declaration above by the candidate is true to the best of my knowledge and that this project report is acceptable for evaluation for the CS5999 PG Diploma Project.

.....

Dr. Uthayasanker Thayasivam

.....

Date

ABSTRACT

Every online product selling applications are having review systems for their customers to review the products that they have purchased. Customers' reviews about the product will be either negative or positive and some reviews will give the meaning explicitly and some reviews will have implicit meaning. Nowadays most of the people do purchasing through online as a result there are thousands of reviews for a single product. On the other hand, these reviews will be useful for other customers to decide whether to purchase the product or not by going through the reviews. Mining implicit features from the customer reviews is a fundamental requirement for extracting customers' opinions and summarizing. This research focuses on extracting implicit features from reviews for opinion mining using a word embedding model. It removes noisy words and learn the model parameters automatically and extract the implicit features from customer reviews. Most of the existing researches have focused on implicit feature extraction from Chinese web reviews and only few attempts are made to extract implicit features from English web reviews. Implicit feature extraction was done through supervised, semi-supervised and unsupervised learning approaches. This research focuses on supervised aspect extraction using deep learning. This research proposes a novel and yet simple CNN model employing two types of pre-trained embeddings for aspect extraction: general-purpose embeddings and domain-specific embeddings associated with a Word Embedding based Correlation (WEC) model by integrating advantages of both the translation model and word embedding to extract implicit features. WEC model can score their correlation score for each word in review and feature. Then the CNN is used to identify the feature where the input for CNN is similarity matrix generated using the correlation scores. CNN gives the matching score of the review feature pair as the output and the review's corresponding feature will be identified from the feature set.

Acknowledgements

I would like to express profound gratitude to my advisor, Dr. Uthayasanker Thayasivam, for his invaluable support by providing relevant knowledge, materials, advice, supervision and useful suggestions throughout this research work. His expertise and continuous guidance enabled me to complete my work successfully.

I would like to thank all my colleagues for their help in finding relevant research material, sharing knowledge and experience and for their encouragement.

I am as ever, especially indebted to my parents for their love and support throughout my life. I also wish to thank my loving friends, who supported me throughout my work. Finally, I wish to express my gratitude to all my colleagues at my work place, for the support given to me to manage my MSc research work.

Table of Contents

DECLARATION.....	ii
ABSTRACT	iii
Acknowledgements	iv
LIST OF FIGURES.....	vii
LIST OF ALGORITHMS.....	vii
LIST OF TABLES.....	viii
1 INTRODUCTION.....	1
1.1 Overview	2
1.2 Problem and motivation.....	2
1.3 Research objectives.....	2
1.4 Organization of the Thesis	3
2 LITERATURE SURVEY.....	4
2.1 Overview	4
2.2 Unsupervised Learning.....	4
2.3. Semi Unsupervised Learning.....	15
2.4. Supervised Learning.....	21
2.5. Summary.....	27
3 METHODOLOGY.....	29
3.1 Overview	29
3.2 Background	29
3.3 Approach	35
3.4 Model	37
3.5. Summary.....	42
4 EXPERIMENTS	43
4.1 Overview	43
4.2 Performance Measures.....	43
4.3 Dataset.....	43

4.4 Hyper-parameter.....	44
4.5 Baseline methods	45
4.6 Summary	47
5 RESULTS AND DISCUSSION	48
5.1 Overview.....	48
5.2 Explicit feature extraction results.....	48
5.3 Implicit Feature extraction results.....	49
5.4 Analysis.....	50
5.5 Summary	52
6 CONCLUSION AND FUTURE WORK.....	53
6.1 Conclusion.....	53
6.2 Future Work.....	53
REFERENCES	54
APPENDIX A.....	57

LIST OF FIGURES

	Page
Figure 1: Framework of hybrid association rule mining for implicit feature identification	5
Figure 2: Schematic diagram of the generative model for a hypothetical Dataset	9
Figure 3: The process of deriving domain vectors for candidate features	12
Figure 4: Similarity co-occurrence matrix	20
Figure 5: Framework of classification-based approach	23
Figure 6: Conceptual block diagram of the proposed system architecture	28
Figure 7: The proposed model for extracting explicit features.....	36
Figure 8: Architecture of WEC + CNN.....	39

LIST OF ALGORITHMS

	Page
Algorithm 2.1: Hybrid association rule mining	6
Algorithm 2.2: Implicit Feature Identification using explicit topic mining model and SVM	17
Algorithm 2.3: Classification based Approach: Implicit Feature Identification	25

LIST OF TABLES

	Page
Table 1: Sample customer reviews on phone	1
Table 2: Hybrid association rule mining: The best performance of using all rules	7
Table 3: Performance comparison of Generative Feature Language Model with baseline methods	10
Table 4: Feature-oriented opinion determination: Best results of extracting domain specific features on D1 – D10	13
Table 5: Rule-Based Approach: Results of experiment on aspect based sentiment analysis data (Semeval 2014)	15
Table 6: Explicit topic mining model: Best performance of different methods	18
Table 7: Opinion Mining Using Clustering: Results of implicit feature identification	21
Table 8: Opinion Mining Using Clustering: Results of comparison	21
Table 9: Classification based Approach: Results of implicit feature identification	25
Table 10: Summary: Methods used to identify implicit features.....	26
Table 11: SemEval 14 and 16 datasets.....	42
Table 12: Human annotated dataset from Liu et al.	42
Table 13: Automatically annotated dataset from Karmaker et al.	42
Table 14: Explicit features F1 score comparison results.....	47
Table 15: Implicit features F1 score, Precision and Recall comparison results.....	48
Table 16: Feature words and their top correlated words.....	49
Table 17: True Positive, False Positive, False Negative and True Negative counts of the evaluation dataset	55