UNIVERSITI PUTRA MALAYSIA

FUZZY SEMANTIC CLASSIFIER FOR DETERMINING STRENGTH LEVELS OF CUSTOMER PRODUCT REVIEWS

SAMANEH NADALI

FSKTM 2012 11

FUZZY SEMANTIC CLASSIFIER FOR DETERMINING STRENGTH LEVELS OF CUSTOMER PRODUCT REVIEWS

By

SAMANEH NADALI

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfillment of the Requirements for the Degree of Master of Science

August 2012

© COPYRIGHT UPM

To my beloved mother and father

© COPYRIGHT UPM

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfillment of the requirements for the degree of Master of Science

FUZZY SEMANTIC CLASSIFIER FOR DETERMINING STRENGTH LEVELS OF CUSTOMER PRODUCT REVIEWS

By

SAMANEH NADALI

August 2012

Chairman: Masrah Azrifah Azmi Murad, PhD

Faculty: Computer Science and Information Technology

Opinion Mining (OM) is one of the new paradigms of information retrieval and computational linguistics. This paradigm is not only concerned with document topic but also the opinion which is expressed. The most challenging area in OM is finding the orientation of customer feeling in reviews such as blogs, product reviews and so on. Opinion about products is nowadays available from blogs and review sites. So, extracting opinion from these reviews help the user as well merchants to track the most suitable choice of product.

There are various tasks in OM. Classification of customer reviews into positive, negative and neutral classes (also known as semantic classification) is one of the tasks

© COPYRIGHT UPM

that help product manufacturers or businesses to easily identify orientation of their product services.

Previous studies focused on the automatic identification of opinion i.e. classifying reviews into positive, negative and neutral only. However, for some applications like flame detection or information analysis, recognizing opinion only might not be sufficient. Thus, identifying strength of opinion is considered as one of the propounded problems from the early days.

In this thesis, we extended the holistic lexicon-based approach to opinion mining presented in (Ding *et al*., 2008), in which the researcher did not focus on finding the strength levels of opinion of each product reviews.

To address the mentioned problem, a Fuzzy Semantic Classifier (FSC) is proposed to identify semantic orientation of customer product reviews at a granularity levels such as *very strong*, *strong*, *moderate*, *weak*, and *very weak* for each positive and negative class by combining opinion words (i.e. adverb, adjective, verb, and noun). We used fuzzy logic as it is not only using non-numerical values but also it introduces the notion of linguistic variables to overcome the uncertainty of natural language.

The proposed classifier (FSC) has been tested on eight benchmark datasets introduced by (Ding *et al*., 2008). The results of the study showed that a Fuzzy Semantic Classifier (FSC) gave various strength of levels classification in customer product reviews which

© COPYRIGHT UPM

leads to multi understandability of customer opinions. The percentage of similarity between FSC and human classifications is 74%. This means that the FSC is able to classify various strength levels to *very strong*, *strong*, *moderate*, *weak* and *very weak* for each positive and negative class similar to human.

© COPYRIGHT UPM

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Master Sains

PENGELAS SEMANTIK SAMAR UNTUK MENENTUKAN TAHAP KEKUATAN BAGI ULASAN PRODUK PELANGGAN

Oleh

SAMANEH NADALI

Ogos 2012

Pengerusi: Masrah Azrifah Azmi Murad, PhD

Fakulti: Sains Komputer dan Teknologi Maklumat

Perlombongan Pendapat (OM) adalah salah satu paradigma baru bagi carian maklumat dan linguistik pengkomputeran. Paradigma ini bukan hanya mempedulikan topic dokumen tetapi juga pendapat yang dinyatakan. Bahagian yang paling mencabar di dalam OM adalah dengan mencari orientasi perasaan pelanggan di dalam ulasan seperti blog, ulasan produk dan sebagainya.

Terdapat pelbagai tugas di dalam OM. Pengelasan ulasan pelanggan ke kelas positif atau negatif (juga dikenali sebagai pengelasan semantik) merupakan salah satu tugas yang dapat membantu pengeluar produk atau perniagaan untuk mudah mengenal pasti orientasi perkhidmatan produk mereka.

Kaedah-kaedah dahulu hanya tertumpu kepada pengenalan automatik untuk pendapat iaitu pengelasan ulasan-ulasan kepada positif dan negatif sahaja. Walau bagaiana pun,

© COPYRIGHT UPM

untuk sebahagian penggunaan seperti pengesanan bara atau analisa maklumat, mengenal pasti pendapat sahaja adalah tidak mencukupi. Dengan itu, mengenal pasti kekuatan pendapat adalah merupakan salah satu masalah yang dikemukakan pada peringkat awal.

Dalam tesis ini kami akan memanjangkan pendekatan holistik yang berpusat leksikon disajikan dalam (Ding *et al*., 2008), di mana penyelidik tidak menumpukan perhatian untuk mencari kekuatan pendapat bagi setiap ulasan produk.

Kami mencadangkan satu pengelas semantik kabur (FSC) untuk mengenal pasti orientasi semantik bagi ulasan produk pelanggan di peringkat granulariti seperti *sangat kuat, kuat, sedang, lemah* dan *sangat lemah* untuk setiap kelas positif dan negative dengan menggabungkan kata-kata pendapat (iaitu kata keterangan, kata sifat, kata kerja dan kata nama). Kami menggunakan logik kabur kerana ia bukan hanya membolehkan penggunaan nilai tidak-berangka tetapi juga memperkenalkan gagasan pembolehubah linguistik untuk mengatasi ketidakpastian bagi bahasa tabii.

Kaedah yang dicadangkan telah diuji pada lapan tanda aras set data diperkenalkan oleh (Ding et al, 2008.). Prestasi kajian menunjukkan bahawa pengelas semantik kabur memberikan pelbagai pengelasan aras kekuatan di dalam ulasan produk pelanggan yang dapat membawa kepada pelbagai pemahaman bagi pendapat pelanggan. Peratus persamaan antara FSC and pengelasan manusia adalah 74%. Ini bermakna FSC boleh mengelaskan aras kekuatan kepada sangat kuat, kuat, sedang, lemah dan sangat lemah untuk setiap kelas positif dan negatif setara dengan manusia.

© COPYRIGHT UPM

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my beloved mother, my supportive father, my brothers and sister who are always there for me. This thesis would not have been possible without their love and understanding. I would like to dedicate this thesis to my parents, for their patients, unconditional love and generosity during my whole life.

My Special thanks to my uncles, Dr Jamshid Mohebalian and Dr Iraj Mohebalian for their impressive help in my thesis.

I would like to thank my compassionate supervisor, Dr Masrah Azrifah Azmi Murad, who always guides me through all ups and downs, joyful and hopeless moments during my research. I truly appreciated her support, concerns, times and sincerity I received during my study.

I would like to thank my co-supervisor, Dr Rabiah Abdul Kadir, for her advices and insightful comments, which guides me through the proper direction. I am indebted for her knowledge and helpful contributions on this thesis.

Finally, thanks God for giving me another opportunity to know myself by living in Malaysia.

© COPYRIGHT UPM

APPROVAL

I certify that an Examination Committee has met on **date of viva** to conduct the final examination of **Samaneh Nadali** on her **Master of Science** thesis entitled " **A FUZZY SEMANTIC CLASSIFIER TO DETERMINE THE STRENGTH LEVELS OF CUSTOMER PRODUCT REVIEWS**" in accordance with Universiti Pertanian Malaysia (Higher Degree) Act 1980 and Universiti Pertanian Malaysia (Higher Degree) Regulations 1981. The Committee recommends that the candidate be awarded the relevant degree. Members of the Examination Committee are as follows:

Chairman, PhD
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Examiner 1, PhD
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Internal Examiner)

Examiner 2, PhD
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(External Examiner)

External Examiner, PhD
Faculty of Science and Technology
(External Examiner)

BUJANG KIM HUAT,PhD
Professor and Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

© COPYRIGHT UPM

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfillment of the requirements for the degree of Master of Science. Members of the Supervisory Committee were as follows:

Masrah Azrifah Azmi Murad, PhD
Senior Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Rabiah Abdul Kadir, PhD
Senior Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

BUJAN BIN KIM HUAT,PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

© COPYRIGHT UPM

DECLARATION

I declare that the thesis is my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously, and is not concurrently, submitted for any other degree at UPM or other institutions.

_____

SAMANEH NADALI

Date:

© COPYRIGHT UPM

TABLE OF CONTENTS

		Page
ABSTRACT		iii
ABSTRAK		vi
ACKNOWLEDGEMENTS		viii
APPROVAL		ix
DECLARATION		xi
LIST OF TABLES		xv
LIST OF FIGURES		xvi
LIST OF ABBREVIATIONS		xviii
CHAPTER		
1	**INTRODUCTION**	**1**
	1.1 Background	1
	1.2 Problem Statement	3
	1.3 Research Objective	4
	1.4 Research Scope	5
	1.5 Research Contribution	7
	1.6 Overview of Thesis	8
2	**LITERATURE REVIEW**	**9**
	2.1 Introduction	9
	2.2 Opinion Mining	9
	2.3 Sentiment and Subjectivity Classification	10
	2.3.1 Document-Level Sentiment Classification	11
	2.3.2 Sentence-Level Sentiment Classification	15
	2.4 Opinion Lexicon Generation	20
	2.4.1 Manual Approach	21
	2.4.2 Dictionary-based Approach	21
	2.4.3 Corpus-based Approach	22
	2.5 Feature-based Sentiment Analysis	25
	2.5.1 Feature Extraction	25
	2.5.2 Opinion Orientation Identification	28
	2.6 Fuzzy Logic	32

© COPYRIGHT UPM

2.6.1 Fuzzy Sets 32
2.6.2 Membership Functions 34
2.6.3 Fuzzy IF-THEN Rules 35
2.6.4 Fuzzy Reasoning 36
2.6.5 Fuzzy Inference Systems 36
2.7 Fuzzy logic in NLP 40
2.8 Fuzzy Logic in Opinion Mining 43
2.9 Summary 45

3 RESEARCH METHODOLOGY 46
3.1 Introduction 46
3.2 STEP 1: Literature Review 48
3.3 STEP 2: Design of Proposed Fuzzy Semantic Classifier (FSC) 49
3.3.1 Product Review 49
3.3.2 Preprocessing 50
3.3.3 Define Opinion Words 51
3.3.4 Opinion Words Orientation 51
3.3.5 Linguistic Rules 51
3.3.6 Fuzzy Logic System Phase 52
3.4 STEP 3: Implementation 54
3.4.1 Datasets 54
3.4.2 Evaluation Metrics 55
3.4.3 Experimental Design 58
3.5 STEP 4: Comparison Results 59
3.6 Summary 60

4 PROPOSED FUZZY SEMANTIC CLASSIFIER (FSC) 61
4.1 Introduction 61
4.2 Formulation of FSC 62
4.3 Extracting Opinion Words 64
4.4 Recognition Opinion Words' Orientation 66
4.5 Pre-Processing 68
4.5.1 Stop-Words Removing 68
4.5.2 Stemming 69
4.6 Design Fuzzy Sentiment Classifier (FSC) 70

© COPYRIGHT UPM

4.7 Fuzzy Semantic Classifier (FSC) 71
4.7.1 Fuzzification 75
4.7.2 Membership Function Designing 76
4.7.3 Fuzzy Rules Designing 78
4.7.4 Defuzzification 81
4.8 FSC Evaluation 82
4.9 Summary 83

5 RESULTS AND DISCUSSION 84
5.1 Introduction 84
5.2 Comparison of the Results between FSC and Previous Method 84
5.3 Questionnaire Survey Results for Fuzzification 85
5.4 Results of Performance Analysis by Humans 88
5.5 Summary 105

6 CONCLUSIONS AND RECOMMENDATION 107
6.1 Introduction 107
6.2 Conclusion 107
6.3 Future work 109

REFERENCES 111
BIODATA OF STUDENT 119
LIST OF PUBLICATIONS 120
APPENDIX A 121
APPENDIX B 123
APPENDIX C 124
APPENDIX D 128
APPENDIX E 138
APPENDIX F 139

© COPYRIGHT UPM