FUZZY BASED IMPLICIT SENTIMENT ANALYSIS ON QUANTITATIVE SENTENCES

AMIR HOSSEIN YAZDAVAR

UNIVERSITI TEKNOLOGI MALAYSIA
FUZZY BASED IMPLICIT SENTIMENT ANALYSIS ON QUANTITATIVE SENTENCES

AMIR HOSSEIN YAZDAVAR

A dissertation submitted in partial fulfillment of the requirement of the award of the degree of
Master of Science (Computer Science)

Faculty of Computing
Universiti Teknologi Malaysia

SEPTEMBER 2013
This dissertation is dedicated to my beloved mother and father for their endless support and encouragement.
ACKNOWLEDGEMENT

Praise be to Allah, the Most Gracious and Most Merciful Who has created the mankind with knowledge and wisdom. First and foremost i would like to express my thanks to Almighty ALLAH on the successful completion of this research work and thesis.

Apart from the efforts of myself, the success of any project depends largely on the encouragement and guidelines of many others. I would like to thank my dear wife Monireh for her personal support, kindness and great patience at all times. My parents have given me their unequivocal support throughout, as always, for which my mere expression of thanks likewise does not suffice.

I would like to show my greatest gratitude to my supervisor Professor Dr. Naomie Salim for her advice and guidance provided throughout my study. Her trust, knowledge and friendly personality have always been an inspiration for me and will deeply influence my career and future life. My gratitude and special thanks is extended to my examiners Dr. Alex Sim Tze Xiang and Dr. Roliana Ibrahim for their helpful suggestions and comments.
ABSTRACT

With the rapid growth of social media on the web, emotional polarity computation has become a flourishing frontier in the text mining community. However, it is challenging to understand the latest trends and summarise the state or general opinions about products due to the big diversity and size of social media data and this creates the need of automated and real time opinion extraction and mining. On the other hand, the bulk of currently research has been devoted to study the subjective sentences which contain opinion keyword and limited work has been reported for objective statement that implies sentiment. In this regard, fuzzy based knowledge engineering model has been developed for sentiment classification of special group of such sentences including the change or deviate from desired range or value. Drug reviews are the rich source of such statements. Therefore, in this research, 210 reviews were collected from patient’s review for building corpus. These reviews have been selected from different cholesterol lowering drugs. Medical experts cooperated in this research for building Gold standard corpus. Pre-processing operations including extracting medical terms and their corresponding values have been done on this corpus. An appropriate technique has been developed to map each of these medical terms to their corresponding values. Resulted documents were stored into XML file. Determining sentiment polarity of each sentence employing fuzzy knowledge based system is the next step of this research. The main conclusion through this study is, in order to increase the accuracy level of drug opinion mining system, objective sentences which imply opinion should be taken into consideration.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>DECLARATION</td>
<td>vii</td>
<td></td>
</tr>
<tr>
<td>DEDICATION</td>
<td>viii</td>
<td></td>
</tr>
<tr>
<td>ACKNOWLEDGEMENT</td>
<td>ix</td>
<td></td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>ABSTRAK</td>
<td>xi</td>
<td></td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>xii</td>
<td></td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>xiii</td>
<td></td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>xiv</td>
<td></td>
</tr>
<tr>
<td>LIST OF ABBREVIATIONS</td>
<td>xv</td>
<td></td>
</tr>
<tr>
<td>LIST OF APPENDICES</td>
<td>xvi</td>
<td></td>
</tr>
</tbody>
</table>

1 INTRODUCTION 1

1.1 Introduction 1
1.2 Problem Background 2
1.3 Problem Statement 4
1.4 Project Aim 5
1.5 Project Objectives 5
1.6 Project Scope 5
1.7 Significance of Project 6
1.8 Organization of the Report 6
1.9 Summary 7

2 LITERATURE REVIEW 8

2.1 Introduction 8
2.2 Sentiment Analysis and Opinion Mining 8
  2.2.1 Document Level Analysis 9
    2.2.1.1 Unsupervised Learning Approach 9
    2.2.1.2 Supervised Method 10
  2.2.2 Sentence Level Analysis 11
    2.2.2.1 Subjectivity Classification 11
    2.2.2.2 Sentence Sentiment Classification 13
  2.2.3 Aspect Based Sentiment Analysis 14
    2.2.3.1 Aspect Extraction 14
    2.2.3.2 Grouping Aspect into Categories 17
    2.2.3.3 Aspect Sentiment Classification 18
    2.2.3.4 Entity, Time, Opinion holder 19

2.3 Biomedical text Mining 19
  2.3.1 Extracting Information in Medical 20
    2.3.1.1 Name Entity Recognition 20
    2.3.1.2 Relation Extraction 23
    2.3.1.3 Event Extraction 23
    2.3.1.4 Knowledge Resources and Tools 24
  2.3.2 Fuzzy Application System in Medical Domain 25

2.4 Sentiment Analysis in Medical Domain 26
  2.4.1 Previous Work and Challenges 26

2.5 Summary 30

3 METHODOLOGY 31
  3.1 Introduction 31
  3.2 Operational Framework 31
    3.2.1 Planning Phase 35
    3.2.2 Collecting Drug Reviews 35
    3.2.3 Preparing Performance Measurements 36
    3.2.4 Analyzing and Categorizing Numerated Sentences 36
3.2.5 Technique Identification 37
  3.2.5.1 GATE Preprocessing Framework 38
3.2.6 Removing Non-numerated Sentences 39
3.2.7 Extracting Quantitative Medical Terms and Values 40
3.2.8 Mapping Values to Terms 40
3.2.9 Storing Opinionated Sentences with their Features 41
3.2.10 Preparing Input Data for Fuzzy System 41
3.2.11 Development of Fuzzy Logic Technique 41
  3.2.11.1 Knowledge Acquisition by Interview and Literature 42
  3.2.11.2 Structure of Proposed Fuzzy Expert System 45
  3.2.11.3 Implementation of Fuzzy Expert System 50
3.2.12 Preparing Data For evaluation 50
3.2.13 Programming of techniques 50
3.2.14 Evaluation 50
3.3 Instrumentation 52
3.4 Summary 53

4 EXPERIMENTAL RESULTS AND DISCUSSION 54
4.1 Introduction 54
4.2 Preprocessing and Intermediate Results and Analysis 54
  4.2.1 Building Corpus Results 55
    4.2.1.1 Corpus Data Distribution 57
    4.2.1.2 Gold Standard Corpus Characteristics 58
  4.2.2 Corpus Text Segmentation Results 60
    4.2.2.1 English Tokeniser Results 61
    4.2.2.2 Sentence Splitter Results 62
    4.2.2.3 ANNI Gazetteer Results 62
    4.2.2.4 JAPE Transducer Results 64
4.2.3 Storing Result of Extraction Phase into XML
4.2.4 Result of Parsing XML File
4.3 Results for Fuzzy Expert System
4.3.1 Results of Fuzzification
4.3.2 Results of Rule Evaluation
4.3.3 Results of Rule Aggregation and Defuzzification
4.4 Results of Preparation XML File for Evaluation
4.5 Overall Evaluation using Precision & Recall and Analysis
4.6 Discussion
4.7 Summary

5 CONCLUSION
5.1 Introduction
5.2 Findings
5.3 Contribution of Study
5.4 Future Work

REFERENCES
Appendices A-G
# LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE NO.</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Accepted ranges of some quantitative medical terms</td>
<td>22</td>
</tr>
<tr>
<td>2.2</td>
<td>Different types of numerated sentences in medical reviews</td>
<td>29</td>
</tr>
<tr>
<td>3.1</td>
<td>Various types of Statin drugs effectiveness for lowering LDL</td>
<td>42</td>
</tr>
<tr>
<td>3.2</td>
<td>Various types of Statin drugs effectiveness for increasing HDL</td>
<td>43</td>
</tr>
<tr>
<td>3.3</td>
<td>Various types of Statin drugs effectiveness for lowering Triglyceride</td>
<td>43</td>
</tr>
<tr>
<td>3.4</td>
<td>Various types of Statin drugs effectiveness for lowering Cholesterol</td>
<td>44</td>
</tr>
<tr>
<td>3.5</td>
<td>Niacin’s effect on various cholesterol items</td>
<td>45</td>
</tr>
<tr>
<td>3.6</td>
<td>List of used Fuzzy sets</td>
<td>47</td>
</tr>
<tr>
<td>3.7</td>
<td>Hardware requirement</td>
<td>52</td>
</tr>
<tr>
<td>3.8</td>
<td>List of used software</td>
<td>53</td>
</tr>
<tr>
<td>4.1</td>
<td>Corpus characteristics</td>
<td>57</td>
</tr>
<tr>
<td>4.2</td>
<td>Polarity distributions over the Gold standard corpus</td>
<td>59</td>
</tr>
<tr>
<td>4.3</td>
<td>Various forms of quantitative medical terms employed by patients</td>
<td>63</td>
</tr>
<tr>
<td>4.4</td>
<td>Extracted Increasing &amp; decreasing verbs</td>
<td>63</td>
</tr>
<tr>
<td>4.5</td>
<td>Confusion matrix for proposed algorithm</td>
<td>72</td>
</tr>
<tr>
<td>4.6</td>
<td>Overall precision and recall</td>
<td>72</td>
</tr>
<tr>
<td>4.7</td>
<td>A sample of false positive &amp; false negative result</td>
<td>75</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE NO.</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Operational Framework</td>
<td>33</td>
</tr>
<tr>
<td>3.2</td>
<td>Diagrammatic summary of research methodology</td>
<td>34</td>
</tr>
<tr>
<td>3.3</td>
<td>Gate preprocessing framework</td>
<td>39</td>
</tr>
<tr>
<td>3.4</td>
<td>Sentiment fuzzy set</td>
<td>49</td>
</tr>
<tr>
<td>4.1</td>
<td>A Sample of collected reviews</td>
<td>55</td>
</tr>
<tr>
<td>4.2</td>
<td>A Sample of stored reviews in XML file</td>
<td>56</td>
</tr>
<tr>
<td>4.3</td>
<td>A Sample of constructed corpus in GATE</td>
<td>56</td>
</tr>
<tr>
<td>4.4</td>
<td>Medical terms distribution over the constructed corpus</td>
<td>57</td>
</tr>
<tr>
<td>4.5</td>
<td>Drug type distributions over the constructed corpus</td>
<td>58</td>
</tr>
<tr>
<td>4.6</td>
<td>A sample of Gold standard preparation process</td>
<td>59</td>
</tr>
<tr>
<td>4.7</td>
<td>Polarity Distributions over Different Types of Sentences</td>
<td>60</td>
</tr>
<tr>
<td>4.8</td>
<td>GATE processing resources</td>
<td>61</td>
</tr>
<tr>
<td>4.9</td>
<td>GATE Tokenization result</td>
<td>61</td>
</tr>
<tr>
<td>4.10</td>
<td>A sample of GATE sentence splitter result</td>
<td>62</td>
</tr>
<tr>
<td>4.11</td>
<td>A sample of GATE Flexible Gazettee result</td>
<td>64</td>
</tr>
<tr>
<td>4.12</td>
<td>A Sample of regular expression used for mapping each medical term to its corresponding values</td>
<td>65</td>
</tr>
<tr>
<td>4.13</td>
<td>A sample of mapping quantitative medical terms to their values result</td>
<td>65</td>
</tr>
</tbody>
</table>
4.14 Distribution of Different Opinionated Sentences Extracted from Corpus 66

4.15 A sample result of extraction phase stored into XML file 67

4.16 A sample of fuzzification step 68

4.17 Result of fuzzification process 69

4.18 A sample of rule evaluation process 69

4.19 A sample of defuzzification process 70

4.20 A sample of annotated sentence with its sentiment stored into XML file 71

4.21 Result of GATE Annotation Diff Tool 72
LIST OF ABBREVIATIONS

NLP - Natural Language Processing
SVM - Support Vector Machines
POS - Part Of Speech
NER - Named Entity Recognition
PMI - Pointwise Mutual Information
KA - Knowledge Acquisition
UMLS - Unified Medical Language System
OOP - Object Oriented Program
GATE - General Architecture for Text Engineering
JAPE - Java Annotation Patterns Engine
HMM - Hidden Markov Model
CRF - Conditional Random Field
API - Application Programming Interface
ANNIE - A Nearly-New Information Extraction
XML - Extensible Markup Language
HMM - Hidden Markov Model
COG - Center Of Gravity
PSO - Particle Swarm Optimization
# LIST OF APPENDICES

<table>
<thead>
<tr>
<th>APPENDIX</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Sample of Corpus Collection Stored in XML file</td>
<td>84</td>
</tr>
<tr>
<td>B</td>
<td>Annotation Guideline</td>
<td>85</td>
</tr>
<tr>
<td>C</td>
<td>Sample of annotated drug reviews by medical experts</td>
<td>86</td>
</tr>
<tr>
<td>D</td>
<td>List of medical terms mapped to appropriate entity</td>
<td>87</td>
</tr>
<tr>
<td>E</td>
<td>Sample of regular expression for mapping medical terms to their corresponding values</td>
<td>88</td>
</tr>
<tr>
<td>F</td>
<td>Sample of generated fuzzy rules</td>
<td>89</td>
</tr>
</tbody>
</table>
1.1 Introduction

Opinion mining or in the other word sentiment analysis, is the study of “what the others think”. Long time before web used as a media for transferring information and opinion, people usually ask each other or their friends to make decision through the various issues such as buying product or planning to vote. However nowadays, there is no limitation to ask others opinion since internet and Web provide a vast pool of reviews from millions of people that we even did not know them.

According to one survey cited by Pang et al. (2008) more than 80 percent of Web users have done at least one online search while purchasing a product. However, monitoring and finding the other’s idea might be confusing and overwhelming since finding relevant sites and reliable opinion through a huge volume of opinionated text in each sites, seems to be impossible. Thus, there is a clear need to build an automatic system to help finding and extracting opinions among different entities. We need to take this into consideration that, sentiment analysis is a Natural Language Processing (NLP) problem. Thus, most of the NLP aspects like negation word, coreference resolution problem and word sense disambiguation are involved in here too. As a result, dealing with this problem, require overcoming a number of challenges. In this regard, one of the basic problems involved in sentiment analysis is considering only sentiment word such as good, wonderful, awful and etc is far from sufficient since one word might have different
orientation in various applications. Furthermore, there are sentences that do not contain any sentiment word whereas they imply opinion. Indeed, these sentences state desirable or undesirable factual information. We believe in order to sentiment analysis system achieve next level of accuracy these sentences should be taken into account.

In the present thesis, we investigate special kind of factual sentences which contain quantitative measurement term that implicitly express sentiment.

1.2 Problem Background

According to (Liu, 2010) opinion can be defined as a quintuple \((e_i, a_{ij}, s_{ijkl}, h_k, t_l)\) where \(e_i\) is the name of entity, \(a_{ij}\) denotes an aspect of entity \(e_i\), \(s_{ijkl}\) shows the sentiment orientation on aspect \(a_{ij}\) of entity \(e_i\), \(h_k\) is opinion holder (the one who express this idea) and \(t_l\) is the time when this idea expressed by its opinion holder. By this definition, sentiment analysis is the task to find all quintuples in the given opinionated document.

Opinions fall within two categories based on the way they are expressed; explicit opinion and implicit opinion. An explicit opinion includes subjective sentences which have opinion keyword. These kinds of statement are easy to detect. The bulk of currently research is devoted to this category.

On the other hand, implicit opinion is an objective statement that implies sentiment. Indeed, they state one desirable or undesirable fact. Limited work has been done in this category. According to Zhang et al., (2011), one special kind of these sentences which is noun and noun phrase that imply opinion has been taken into consideration. However, there exist other approaches which remained unexplored. One group of such approaches is related to change or deviate from desired range or value. Regarding this issue in some application domains, the values of an item have specific properties which denote the sentiment. Indeed, change of
these values to the normal and optimal interval or deviation from the norm range might express positive or negative opinion respectively. For example let the optimal value for total Cholesterol have been defined below than 200 mg/dl, then the sentence “this drug lowered my Cholesterol form 300 to 190” bear positive sentiment since the Cholesterol decreased into optimal value. And in sentence “don’t take this drug, it puts my Blood pressure into 18” implicit opinion by changing from normal range has been expressed.

Furthermore, based on our observation, significant changes might also denote sentiment even the new value would not place in a normal range. For instance in the sentence “it dropped my Cholesterol level from 580 into 250”, although the second value of Cholesterol would not place in the optimal value (below than 200mg/dl) but the sentence express positive sentiment. On the other hand, the sentence “it increased my Cholesterol level from 250 into 580” denotes negative sentiment polarity, thus it is important to consider change direction.

It is worth mentioning to recall that, among these kinds of sentences, there exist some kinds of sentences that do not express any sentiment while containing changes in numeric values. As an example “my doctor changes the normal dosage of Welchol from 624mg to 300mg” is a factual sentence that should be grouped into non-opinionated sentence. In addition, there are many numeric fields which do not show any sentiment and need to be filtered e.g. “Have never felt so bad, like a 100 year old woman (I am 63)”.

In the present thesis, we investigate numerated sentences, not only to categorize them into opinionated and non-opinionated but also to determine whether they contain positive or negative sentiment polarity.
1.3 Problem Statement

Determining sentiment orientation from sentences have been studied by many researchers, considering different method to handle this problem although a large portion of their effort have been devoted to subjective statement that contain sentiment word, objective sentences which contain desirable or undesirable fact can imply sentiment too. One group of these statements that remains unexplored to date, contains numeric values denote change and deviation from normal range. Drug reviews are the rich source of such statement. In this regard, one special characteristic of such sentences is related to certain degree of uncertainty and imprecision involved in them. In the light of our observation, this uncertainty can be regarded from different point of view. First, a large portion of quantitative medical term associated with predefined ranges which try to identify a patient’s status e.g. for Total Cholesterol, below than 200mg/dl considered as desirable, between 200 and 239 is borderline high and greater than 240 defined as high. Thus, these terms can be accepted as fuzzy.

Second, changes in the value of these quantitative terms, might demonstrate improvement, stable condition, and exacerbation of a patient, regardless of where the second value placed. For example, the sentence “my total Cholesterol dropped from 580 to 240” shows improvement in total Cholesterol while the second value of it (240) placed into high range. Intuitively, changes can be grouped into slight, medium and high increased or decreased which can be denoted by fuzzy set theory.

Furthermore, patient’s sentiment might be regarded as positive, neutral and negative based on the factors that have been stated, second numeric value and changes, side effect, opinion words and etc.

Consequently, fuzzy logic is an ideal choice to deal with this problem. Therefore, the present study tries to determine sentiment polarity of numerated sentence by employing fuzzy set theory. There are four main issues which can achieve the goal of this study; what is the method to extract numeric variable in a sentence? How this numbers related to their specific entities? Which numerated
sentence bear a sentiment? And what is the sentiment orientation of opinionated sentence?

1.4 Project Aim

The work aims to develop fuzzy algorithm to classify numerated factual statement contain implicit opinion.

1.5 Project Objectives

This study aim to accomplish the following objectives;

I. To extract quantitative aspects and their corresponding values.
II. To determine opinionated sentences through objective sentences that contained numerated facts.
III. To develop fuzzy rule based decision system for the purpose of identifying sentiment orientation of numerated opinionated objective sentences that contain implicit opinion.
IV. To evaluate suggested algorithm using precision and recall measures.

1.6 Project Scope

I. The corpus collection will consist from 210 drug reviews collected from website: www.askapatient.com
II. 110 reviews will be used for generating fuzzy rules by expert.
III. 100 reviews will be utilized for building test set for determining sentiment orientation.

IV. The corpus collection will be stored in XML file.

V. GATE will be employed to annotate numbers and extract medical terms.

VI. This thesis will apply GATE JAPE Grammar for mapping extracted quantitative medical terms to their associated values.

VII. Fuzzy rules will be formulated from doctors for inference step.

VIII. Fuzzy expert system approach will be taken to accomplish sentiment classification.

1.7 Significance of Project

This study represents a positive step toward implicit opinion mining to achieve higher accuracy system by exploiting fuzzy set theory and natural language processing techniques.

1.8 Organization of the Report

This study is organized into five chapters. Chapter 1 represents the introduction of the study, problem background, project aim, scope, objectives and the significance of the project. Chapter 2 discusses the previous work and the literature review. Chapter 3 explains the methodology employed in this thesis. Chapter 4 represent the experimental results and finally conclusion and findings are shown in Chapter 5.
1.9 Summary

In conclusion, by rapid growth of social media on the web, we confront a huge volume of opinionated documents such as reviews, forum, discussion, blogs and twitters which have been attracted many researchers due to its integral role in not only Natural Language Processing, but its effect on medical, political and social science. However, there are many challenges which have not been solved and require more research. In this study, we deal with one of this problem to increase the accuracy level of existing system.


Hai, Z., K. Chang and J. Kim (2011). Implicit feature identification via co-
ocurrence association rule mining. *Computational Linguistics and Intelligent
Text Processing*: 393-404.

of the tenth ACM SIGKDD international conference on Knowledge discovery
and data mining*, ACM.

Jakob, N. and I. Gurevych (2010). Extracting opinion targets in a single-and cross-
domain setting with conditional random fields. *Proceedings of the 2010
Conference on Empirical Methods in Natural Language Processing*,
Association for Computational Linguistics.

framework for web opinion mining. *Proceedings of the 26th Annual

Jones, P. H., M. H. Davidson, E. A. Stein, H. E. Bays, J. M. McKenney, E. Miller, V.
A. Cain and J. W. Blasetto (2003). Comparison of the efficacy and safety of
rosuvastatin versus atorvastatin, simvastatin, and pravastatin across doses

Processing*: 627-666.

*Mining Text Data*: 415-463.

subjectivity summarization based on minimum cuts. *Proceedings of the 42nd
Annual Meeting on Association for Computational Linguistics*, Association
for Computational Linguistics.


information on pacemaker implantation procedures from clinical reports.
*AMIA Annual Symposium Proceedings*, American Medical Informatics
Association.


Simpson, M. S. and D. Demner-Fushman (2012). Biomedical text mining: A survey

Swaminathan, R., A. Sharma and H. Yang (2010). Opinion mining for biomedical text data: Feature space design and feature selection. *the Nineth International Workshop on Data Mining in Bioinformatics (BIOKDD 2010)*.


