SIDE EFFECTS RECOGNITION AS IMPLICIT OPINION WORDS IN DRUG REVIEWS

MONIREH EBRAHIMI

UNIVERSITI TEKNOLOGI MALAYSIA

SIDE EFFECTS RECOGNITION AS IMPLICIT OPINION WORDS IN DRUG REVIEWS

MONIREH EBRAHIMI

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

AUGUST 2013

This dissertation is dedicated to my beloved mother and father for their endless support and encouragement

ACKNOWLEDGEMENT

Apart from the efforts of myself, the success of any project depends largely on the encouragement and guidelines of many others. I take this opportunity to express my deepest appreciation to all those who provided me the possibility to complete this report.

Above all, I would like to thank my dear husband Amir Hossein for his personal support, love, kindness and great patience at all times. My parents, brothers and sister 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 the continuous support of my research, for her patience, trust, motivation, enthusiasm, and immense knowledge. Without her guidance and persistent help this dissertation would not have been possible. 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

Many opinion mining systems and tools have been developed to provide the user with the attitude of people toward entities and their attribute or the overall polarity of document. Unlike explicit opinion mining limited work has been done on implicit one. Similarly, few works has been done for opinion mining in medical domain whereas it is a domain dependent task especially about implicit opinions. Besides, side effects are one of critical measures to evaluate the patient's opinion about one drug. However, side effect recognition is challenging task since side effects coincide with disease symptoms lexically and syntactically. In this regard, this study tries to extract drug side effects from drug reviews as an integrable implicit opinion word detection algorithm to a medical opinion mining system using rule based and SVM algorithm. Developing each of these techniques requires different preprocessing steps including corpus text segmentation, mapping medical terms to concepts, trigger terms list construction and SVM feature extraction. Also, due to the novelty of this issue, corpus construction carried out. The corpus used in this study has 225 drug reviews manually annotated by a medication expert as a reference standard. After corpus preprocessing, two proposed techniques has been run. In rule based algorithm, regular expressions and trigger terms list has been used to detect drug adverse side effects and discriminate them from disease symptoms. In the other hand, combination of lexical, syntactical, contextual and semantic features leads to the best results in SVM technique. The results show that SVM significantly performs better than rule based algorithm. However, the results of both algorithms are encouraging and a good foundation for future researches. Obviating the limitations and exploiting combined approaches would improve the results.

ABSTRAK

Banyak sistem perlombongan pendapat dan alat telah dibangunkan untuk menyediakan pengguna dengan sikap orang ke arah entiti dan sifat mereka atau kutub keseluruhan dokumen. Berbeza dengan pendapat yang jelas kerja-kerja perlombongan terhad telah dilakukan pada satu tersirat. Begitu juga, beberapa kerja-kerja yang telah dilakukan untuk perlombongan pendapat dalam domain perubatan sedangkan ia merupakan satu tugas yang domain bergantung terutama tentang pendapat tersirat. Selain itu, kesan-kesan sampingan adalah salah satu daripada langkah-langkah kritikal untuk menilai pendapat pesakit kira-kira satu dadah. Bagaimanapun, bahagian pengiktirafan kesan tugas yang mencabar kerana kesan sampingan yang bertepatan dengan gejala-gejala penyakit leksikal dan sintaksis. Dalam hal ini, kajian ini cuba untuk mendapatkan kesan sampingan dadah daripada ulasan dadah sebagai terkamir tersirat pendapat perkataan algoritma pengesanan kepada sistem perlombongan pendapat perubatan menggunakan kaedah berdasarkan dan algoritma SVM. Membangunkan setiap teknik memerlukan langkah-langkah pra pemproses yang berbeza termasuk corpus pembahagian teks, pemetaan segi perubatan kepada konsep, mencetuskan segi senarai pembinaan dan SVM ciri pengekstrakan. Juga, disebabkan oleh sesuatu yang baru isu ini, corpus pembinaan dijalankan. The corpus yang digunakan dalam kajian ini mempunyai 225 ulasan dadah manual dijelaskan oleh pakar perubatan sebagai standard rujukan. Selepas pra pemproses corpus, dua teknik yang dicadangkan telah dijalankan. Dalam kaedah berasaskan algoritma, ungkapan biasa dan mencetuskan senarai istilah telah digunakan untuk mengesan dadah kesan sampingan yang buruk dan membezakan mereka daripada gejala-gejala penyakit. Di sisi lain, gabungan leksikal, sintaksis, ciri-ciri konteks dan semantik membawa kepada hasil yang terbaik dalam teknik SVM. Keputusan menunjukkan bahawa SVM ketara melakukan lebih baik daripada algoritma berasaskan peraturan. Walau bagaimanapun, keputusan kedua-dua algoritma yang menggalakkan dan asas yang baik untuk kajian masa depan. Obviating batasan dan mengeksploitasi pendekatan gabungan akan meningkatkan hasil.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

NLP	-	Natural Language Processing
MRS	-	Manifestation Related Symptom
ADE	-	Adverse Drug Event
SVM	-	Support Vector Machines
POS	-	Part Of Speech
NER	-	Named Entity Recognition
UMLS	-	Unified Medical Language System
PMI	-	Pointwise Mutual Information
NLM	-	National Library of Medicine
EHR	-	Electronic Health Record
WHO	-	World Health Organization
NDFRT	-	National Drug Formulary Reference Terminology
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
ME	-	Maximum Entropy
EMR	-	Electronic Medical Record

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Decision making is the first and important stage of almost all human activities. Human decisions are result of their view point and evaluation of the situation and this evaluation to considerable degree is shaped based on the other's people opinion about the conditions. The explosive growth of social media in the World Wide Web, change dramatically the people's method to express their opinion and consequently their decision making in their life.

Medical domain is not exclusive of this fact. With the vast amount of medical online information and rapid growth of social media in this field, people no longer use the drug before going to the Internet. They first search through bulk of information to find "What other patients/physicians think about this drug?". This is just an example of many applications of online opinion in people's today life.

However, extracting and analyzing opinions manually from the huge volume of texts is a formidable or even impossible task. The aim of automatic sentiment analysis (opinion mining) system is to provide the user with the attitude of people toward entities and their attribute or the overall polarity of a document. It is an active research area in natural language processing and text mining which involves several subproblems. Unlike implicit opinion mining, a significant amount of research has been done on explicit opinion mining. Similarly limited work has been done for

opinion mining in medical domain whereas it is a domain dependent task especially about implicit opinions. Indeed, medical domain texts including drug reviews are full of medical terms and side effects construct large portion of these terms. Side effects can imply both positive and negative opinion about one drug. Even so, talking about drug side effect is rarely positive and positive terms are more related to drug effectiveness.

In this thesis we develop a system to detect side effects in the drug reviews as a subtask of detecting implicit opinions in medical sources. We will also propose and compare the regular based and classification approaches for recognizing adverse side effects and discriminating them from disease symptoms.

1.2 Problem Background

The reason which provides strong motivation for research on this problem is three-fold. The first angle is opinion mining previous works; due to this fact that this problem is defined as a subtask for an opinion mining system. The main goal of a whole opinion mining system is achieving to a high precision and recall in all of its tasks. Opinion words play an integral role in these systems. Much of the work to date zero in on extracting explicit opinions. However, to reach a high recall considering implicit opinions is essential. So far, minimal work has been done in implicit opinions and (Zhang & Liu, 2011a) is one of the few attempts to deal with this problem. They propose statistical method to extract some noun product features that imply opinion in four data sets including drug reviews. However, this work gives a general approach for all domains whereas opinion mining (especially implicit opinion mining) is a domain and context dependent task and to reach good and applicable results it would be better to exploit some domain knowledge and features.

The second motivation is an importance of opinion mining on drug reviews and a few accomplished works on it. Yalamanchi (2011) in his thesis develop a system called Sideffective to crawl, rank and analyze patient's sentiment about

medications from the side effect point of view. They extract the side effects without using a precompiled list; search and rank the negativity of reviews with respect to side effects using their proposed Negativity Meter formula. They apply a lexicon-based approach which considers side effect as a negative implicit opinion for medicine but one of the main possible shortcomings of this approach is related to its side effect extraction algorithm that is proposed by Rajagopalan (2011). In this algorithm, they filter the most frequent side effects of one drug in order to remove the disease symptom from the list of side effects. The hypothesis behind this algorithm is that top side effect of one drug is a disease symptom that this drug will cure. Although this assumption may be true in some cases, but the algorithm may eliminate the top most important side effect of one drug in other cases.

The third motivation is related to the studies on side effect extraction and differentiating between disease symptom and drug adverse reaction because of their profound impact on the accuracy of implicit opinion detection. In this research direction, Weeber *et al.* (2000) suggested a text based discovery system (DAD) which was a concept-based NLP system for PubMed citations. He claimed that this system has ability to discover new hidden association between adverse drug reaction and a drug. Wang *et al.* (2010) took a different approach using statistical association between extracted entities. They use co-occurrence criteria to detect the type of relation between entities and use the section name where the entities occur to increase the accuracy. In another work, Li (2011) extract side effects from online patient drug reviews on a special class of drugs using statistical and parsing techniques.

In summary, despite the importance of implicit opinion mining in medical domain texts and the essential role of powerful side effect extraction algorithm in achieving to this goal, yet there is possibly no such precise algorithm to detect drug side effects in drug reviews. In fact, all of these existing algorithms consider disease symptom as a drug side effect or focus on only one drug or are not capable to extract all side effects. Therefore, using combined approach and considering the context is essential in this regard. In this project, we will extract side effects from drug reviews by detect and eliminating the disease symptom from the result list to increase the

accuracy of this algorithm. This algorithm can integrate as an implicit opinion word detection part into medical opinion mining system.

1.3 Problem Statement

One issue that remains almost unexplored in medical opinion mining is detecting implicit sentiment words and phrases which shows desirable or undesirable facts about one medicine in the drug review. One group of these words and phrases are drug side effects which cover large portion of drug reviews. On the other hand, most of drug reviews are narrative and contain experiences of patients. So just detecting these kinds of words is not solution to this problem. For instance some sentences shows the general status of patients or their symptom before taking the drug or many other situations that we should take into account. As an example in the sentence “I take Alprazolam when I’m having anxieties about whatever it may be”, “anxiety” is a disease symptom and the reason for taking this psychoactive drug by the patient. Indeed, in this sentence the anxiety does not imply any opinion about the related drug. In contrast, in the sentence “Two month into taking this pill, I started having severe anxiety and anxiety attacks” which is part of a review on one birth control drug, anxiety is a drug side effect and is used to show the negative opinion of patient about the side effects of this drug. Therefore, despite the nonexistence of any opinion word, this sentence is opinionated and a drug side effect should be considered as a negative implicit opinion word in this context.

This thesis try to solve this problem by considering the context in which the medical concept occurs to differentiate between disease-manifestation related symptom (MRS) and drug-adverse drug event (ADE) and extract the second one as sub problem in a medical opinion mining.

To address this issue, we propose two approaches. In the first one, we have refined the idea of using regular expression for identifying contextual features from clinical text suggested by Chapman *et al.* (2007).The second method is machine

learning approach using support vector machines (SVM). Finally we will compare the results of two approaches from the precision and recall point of view.

There are two issues that can achieve the goal of this study: How can we detect the medical concepts in the drug review that imply an opinion on the drug? Which one is the better scheme for drug side effect extraction: regular expression or machine learning approach?

1.4 Project Aim

The study aims to investigate the use of regular expression based and machine learning (SVM) techniques for extracting drug side effects from drug reviews as an integrable implicit opinion word detection algorithm to a medical opinion mining system. The result of these two proposed algorithms will be compared to achieve a side effect detection mechanism that guarantee tagging of minimal disease symptom or other medical terms as a side effect (high precision) while insuring drug side effects are not overlooked (high recall) for effective recognition.

1.5 Project Objectives

Specific objectives of this project are:

1. To develop regular expression based approach for the purpose of detecting drug side effects.
2. To develop machine learning approach for the same purpose.
3. To compare both techniques for effective and efficient detection of side effects.

1.6 Project Scope

1. The corpus collection will contain 225 manually annotated drug reviews collected from DrugRatingz.com.
2. Two techniques will be developed: regular expression and machine learning based.
3. 70 reviews will be used as a development set and 155 as a test set for regular expression based scheme.
4. Support vector machines (SVM) will be exploited for machine learning based scheme.
5. 5-fold cross-validation will be used to perform the evaluation of the SVM model.
6. The MetaMap program will accomplish mapping of medical terms in the drug reviews to UMLS concepts.
7. Trigger terms list will be constructed for this project.

1.7 Significance of Project

The result of this study when integrate to a medical opinion mining system can help not only the patients to assess the drug before taking it but also physicians and drug producer organizations to consider user's feedback in their decision making. This algorithm can also be used in pharmacovigilance systems.

1.8 Organization of the Report

The thesis is organized as follow. Chapter 2 discusses the previous works and literature review on existing techniques in opinion mining and in medical domain.

Chapter 3 explains the methodology of the project. Chapter 4 analyses the experimental results. Finally, the conclusion of this study is given in chapter 5.

REFERENCES

Ananiadou, S., S. Pyysalo, J. i. Tsujii and D. B. Kell (2010). Event extraction for systems biology by text mining the literature. *Trends in biotechnology* 28(7): 381-390.

Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings of the AMIA Symposium*, American Medical Informatics Association.

Cao, H., G. Hripcsak and M. Markatou (2007). A statistical methodology for analyzing co-occurrence data from a large sample. *Journal of biomedical informatics* 40(3): 343-352.

Chapman, W. W., D. Chu and J. N. Dowling (2007). ConText: An algorithm for identifying contextual features from clinical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, Association for Computational Linguistics.

Cohen, A. M. and W. R. Hersh (2005). A survey of current work in biomedical text mining. *Briefings in bioinformatics* 6(1): 57-71.

Cunningham, H., D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts and D. Damljanovic (2011). Developing language processing components with gate version 6 (a user guide). *Internet:* http://gate. *ac. uk/sale/tao/tao. pdf; last accessed= 12.05*.

Denecke, K. and W. Nejdl (2009). How valuable is medical social media data? Content analysis of the medical web. *Information Sciences* 179(12): 1870-1880.

Goeuriot, L., J.-C. Na, W. Y. M. Kyaing, S. Foo, C. Khoo, Y.-L. Theng and Y.-K. Chang (2011). Textual and informational characteristics of health-related social media content: A study of drug review forums. *Proceedings of the Asia-Pacific Conference on Library and Information Education and Practice: Issues, Challenges and Opportunities*.

Gooch, P. and A. Roudsari (2011). A tool for enhancing MetaMap performance when annotating clinical guideline documents with UMLS concepts.

Li, Y. A. (2011). *Medical data mining: Improving information accessibility using online patient drug reviews*. Massachusetts Institute of Technology

Liu, B. (2007). *Web data mining: exploring hyperlinks, contents, and usage data*. Springer.

Liu, B. and L. Zhang (2012). A survey of opinion mining and sentiment analysis. *Mining Text Data*: 415-463.

Narayanan, R., B. Liu and A. Choudhary (2009). Sentiment analysis of conditional sentences. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, Association for Computational Linguistics.

Niu, Y., X. Zhu, J. Li and G. Hirst (2005). Analysis of polarity information in medical text. *AMIA Annual Symposium Proceedings*, American Medical Informatics Association.

Rajagopalan, S. (2011). *Sideffective-system to mine patient reviews: side effect extraction*. Rutgers, The State University of New Jersey

Simpson, M. S. and D. Demner-Fushman (2012a). Biomedical text mining: A survey of recent progress.*Mining Text Data* 465-517, Springer.

Simpson, M. S. and D. Demner-Fushman (2012b). Biomedical text mining: A survey of recent progress. *Mining Text Data*: 465-517.

Skentzos, S., M. Shubina, J. Plutzky and A. Turchin (2011). Structured vs. unstructured: factors affecting adverse drug reaction documentation in an

EMR repository. *AMIA Annual Symposium Proceedings*, American Medical Informatics Association.

Spasić, I., F. Sarafraz, J. A. Keane and G. Nenadić (2010). Medication information extraction with linguistic pattern matching and semantic rules. *Journal of the American Medical Informatics Association* 17(5): 532-535.

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)*.

Wang, X., H. Chase, M. Markatou, G. Hripcsak and C. Friedman (2010). Selecting information in electronic health records for knowledge acquisition. *Journal of biomedical informatics* 43(4): 595-601.

Weeber, M., H. Klein, A. R. Aronson, J. G. Mork, L. De Jong-van Den Berg and R. Vos (2000). Text-based discovery in biomedicine: the architecture of the DAD-system. *Proceedings of the AMIA Symposium*, American Medical Informatics Association.

Yalamanchi, D. (2011). *Sideffective-system to mine patient reviews: sentiment analysis*. Rutgers University-Graduate School-New Brunswick

Zhang, L. and B. Liu (2011a). Extracting Resource Terms for Sentiment Analysis. *IJCNLP*.

Zhang, L. and B. Liu (2011b). Identifying Noun Product Features that Imply Opinions. *ACL (Short Papers)*.

Zweigenbaum, P., D. Demner-Fushman, H. Yu and K. B. Cohen (2007). Frontiers of biomedical text mining: current progress. *Briefings in bioinformatics* 8(5): 358-375.