香港浸會大學
HONG KONG BAPTIST UNIVERSITY

HKBU Scholars

MASTER'S THESIS

Computational models for mining online drug reviews

Tang, Chao

Date of Award:
2014

Link to publication

General rights
Copyright and intellectual property rights for the publications made accessible in HKBU Scholars are retained by the authors and/or other copyright owners. In addition to the restrictions prescribed by the Copyright Ordinance of Hong Kong, all users and readers must also observe the following terms of use:

- Users may download and print one copy of any publication from HKBU Scholars for the purpose of private study or research
- Users cannot further distribute the material or use it for any profit-making activity or commercial gain
- To share publications in HKBU Scholars with others, users are welcome to freely distribute the permanent URL assigned to the publication

Download date: 03 Feb, 2023

Abstract

Healthcare social media is emerging in recent years with increasing attention on people's health. Online review websites are not only diversified with medicine, hospitals, or doctors but abundant in amount. To discover knowledge from these online reviews, several computational models are proposed.

Online healthcare review websites are facing challenges in conflict of interests among various healthcare stakeholders. To avoid legal complaints and better sustain under such circumstance, we propose a decoupling approach for designing healthcare review websites. Objective components such as medical condition and treatment are remained as the primary parts, as they are generic, impersonal and directly related to patients themselves. Subjective components, however, such as comments to doctors or hospitals are decoupled as secondary parts for sensitive and controversial information and are optional to reviewers. Our proposed approach shows better flexibility in managing of contents in different levels of details and ability of balancing the right of expression of reviewers with other stakeholders.

To identity the patient-reported adverse reactions in drug reviews, we propose a consumer-oriented coding scheme using wordnet synonym and derivational related form. Significant discrepancy of incidences of adverse reactions is discovered between online reviews and clinical trials. We proposed an adverse reaction report ratio model for integrated interpretation of adverse reactions reported in online reviews versus those from clinical trial. Our estimation on average adverse reactions shows high correlation with drug acceptability score obtained from a large-scale meta-analysis.

To investigate the impact of key adverse reactions in patients' perspective, we propose a topic model named Fisher's Linear Discriminant Analysis Projected Non-negative Matrix Factorization (FLDA-projected-NMF) for discovering discriminative features and topics with additional class information. With satisfaction scores

provided in the reviews, discriminative features and topics on satisfaction are discovered and polarities of adverse reactions are estimated based on the discriminative feature weights. Discriminative features and topics on medication duration and on age group are obtained as well. Our method outperforms other supervised methods in evaluation of topic sentiment score and topic interpretation measured by entropy. Patient-reported adverse reaction terms are mined from reviews with comment class label. Some new adverse reactions in depression drug and statin drug are also discovered.

To further study patients' behaviors, we use structural equation modeling for studying the relationship of factors in patients' treatment experience with patients' quality of life. In covariance model, most adverse reactions are found of small covariance except nausea, headache and dizziness. In measurement model, coefficients of individual adverse reactions on latent adverse reaction are correlated to the incidence of adverse reactions. In structural model, we model the relationship of latent adverse reaction, rating score, positive sentiment and negative sentiment. Comparison between the measurement models of rating scores of depression drug and statin drug shows that there could be latent factors to account for the variances of latent rating, which shows correlations with the severity of adverse reactions.

Acknowledgements

First and foremost, I would like to express my deepest gratitude and appreciation to my principal supervisors, Dr. Chun-Hung Li and Dr. William Cheung, for their guidance, encouragement and support throughout my study these years. It is Dr. Li who brought me into this exciting research area and supervised me doing independent research. I am very grateful and indebted to him for his valuable time for discussion with me.

I wish to express my sincere thanks to my co-supervisor Prof. Jiming Liu for his useful advice and encouragement on my postgraduate study. I am grateful to all the faculty members in the department for their academic guidance. I also would like to thank my examination committee and my external examiner.

I want to thank all my friends, my colleagues, especially Dr. Victor Cheng, and the staff in our department for their technical assistance and support.

In this special moment, I would like to express my deepest thanks to my parents for their endless love and unceasing encouragement and support.

Contents

Declaration i

Abstract ii

Acknowledgements iv

Table of Contents vii

List of Tables x

List of Figures xii

1 Introduction 1

2 Challenges in Disseminating Healthcare Knowledge via Social Media 5

2.1 Introduction 5

2.2 Current social healthcare system 7

2.3 A decoupling approach for designing social healthcare system . . . 11

2.4 Discussion 14

2.5 Conclusion 16

3 Integrating Adverse Reaction Reporting in Online Reviews with Clinical Trial 18

3.1 Introduction 18

3.2 Identification of adverse reactions in online reviews 19

3.2.1 Common adverse reactions and clinical trial incidences . . . 19

3.2.2 Online drug reviews acquisition and preprocessing 20

3.2.3 Consumer-oriented coding scheme for adverse reactions . . 21

3.2.4 Measuring the adverse reaction report ratios 22

3.3 Results . . . 27
3.3.1 Adverse reaction incidence of drug reviews . . . 27
3.3.2 Adverse reaction average report ratio and adverse reaction incidence estimation . . . 27
3.3.3 Relationship of adverse reaction and acceptability . . . 28
3.4 Limitations . . . 31
3.5 Compare with prior work . . . 34
3.6 Future work . . . 36
3.7 Conclusion . . . 37

4 Discovering Discriminative Features and Topics with Additional Class Information 38
4.1 Introduction . . . 38
4.1.1 Related work . . . 38
4.1.2 Objectives . . . 39
4.2 Methods . . . 40
4.2.1 Fisher's linear discriminant analysis (FLDA) . . . 41
4.2.2 Non-negative matrix factorization (NMF) . . . 42
4.2.3 FLDA-projected-NMF . . . 42
4.3 Experiments and results . . . 44
4.3.1 Finding discriminative features and topics on satisfaction . . . 46
4.3.2 Extraction of adverse reactions from online reviews . . . 47
4.3.3 Discovery of polarity of adverse reaction . . . 52
4.3.4 Application on gender, age group, and medication duration . . . 54
4.4 Evaluation and discussion . . . 57
4.4.1 Feature selection comparison . . . 57
4.4.2 Comparison with other regression models . . . 60
4.4.3 Topics results comparison . . . 61
4.5 Conclusion . . . 82

5 Mining Online Drug Reviews with Structural Equation Modeling 84
5.1 Introduction . . . 84
5.1.1 Literature review . . . 84
5.1.2 Objective . . . 85

5.2 Data 86
5.3 Methods 88
5.4 Models 92
5.4.1 Covariance model 92
5.4.2 Measurement model 92
5.4.3 Structural model 94
5.5 Results 95
5.5.1 Covariance model 95
5.5.2 Measurement model 97
5.5.3 Structural model 101
5.6 Model evaluation 103
5.7 Discussion 104
5.7.1 Limitation 106
5.8 Conclusion 108

6 Conclusion and Future Work **116**

Bibliography **126**

Curriculum Vitae **127**