UTS UNIVERSITY OF TECHNOLOGY SYDNEY

Reliable Sentiment Analysis in Social Media

by Qian Li

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Qiang Wu, Jian Zhang

University of Technology Sydney
Faculty of Engineering and Information Technology

July, 2020

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, *Qian Li* declare that this thesis, submitted in partial fulfilment of the requirements for the award of Doctor of Philosophy, in thedeclare that this thesis, is submitted in fulfilment of the requirements for the award of *Doctor of Philosophy*,in the *Faculty of Engineering and Information Technology* at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution. This research is supported by the Australian Government Research Training Program.

Production Note:

SIGNATURE: Signature removed prior to publication.

Qian Li

DATE: 29 July, 2020

Abstract

Sentiment analysis in social media is critical yet challenging because the source materials (i.e., reviews posted in social media) are with high complexity, low quality, and uncertain credibility. For example, words and sentences in a textual review may couple with each other, and they may have heterogeneous meanings under different contexts or in different language locales. These couplings and heterogeneities essentially determine the sentiment polarity of the review but are too complex to be captured and modeled. Also, social reviews contain a large number of informal words and typos (a.k.a., noise) but a rare number of vocabularies (a.k.a., sparsity). As a result, most of the existing natural language processing (NLP) methods may fail to represent social reviews effectively. Furthermore, a large proportion of social reviews are posted by fraudsters. These fraud reviews manipulate social opinion, and thus, they disturb sentiment analysis.

This research focuses on reliable sentiment analysis in social media. It systematically investigates the sentiment analysis techniques to tackle three major challenges in social media: high data complexity, low data quality, and uncertain credibility. Specifically, this research focuses on two research problems: general sentiment analysis in social media and fraudulent sentiment analysis in social media. The general sentiment analysis targets on tackling high data complexity and low-quality of social articles that are credible. The fraudulent sentiment analysis handles the uncertain credibility issue, which is common and profoundly affects the precise sentiment analysis in social media. Based on these investigations, this research proposes a serial of methods to achieve reliable sentiment analysis: It studies the polarity-shift characteristics and non-IID characteristics in general paragraphs to capture the sentiment more accurately. It further models multi-granularity noise and sparsity in short text, which is the most common data in social media, for robust short text sentiment analysis. Finally, it tackles the uncertain credibility problem in social media by studying fraudulent sentiment analysis in both supervised and unsupervised scenarios.

This research evaluates the performance and properties of the proposed reliable sentiment analysis methods by extensive experiments on large real-world data sets. It demonstrates that the proposed methods are superior and reliable in social media sentiment analysis.

Dedication

To my son Jewell Zhu...

ACKNOWLEDGMENTS

First of all, I want to thank all the contributions my supervisor Prof. Qiang Wu make during my Ph.D experience. I appreciate it very much for his time, ideas, discussions, and paper polishing to make my Ph.D. studies impressive and stimulating. I am deeply influenced by his professionalism, profound ideas and hard work. Also, his optimism attitude and enthusiasm were contagious and inspiring for me, even during the difficult times in the pursuit of my doctoral degree.

I am also greatly indebted to Prof. Jian Zhang, my co-supervisor, whose constant suggestions, incisive comments and encouragements helped me a lot during my Ph.D. studies.

My sincere thanks also go to all of my colleagues in the Global Big Data Technologies Centre, for their enthusiastic help: Dr.Peng Zhang, Dr. Zongjian Zhang, Xunxiang Yao, Dr. Xiaoshui Huang, Yan Huang, Lu Zhang, Yongshun Gong, Xiaofei Liu, Anan Du, Lingxiang Yao. I appreciate our intensive discussions, in-depth collaboration, and all the fun we have had in the last four years.

I thank the colleagues in the student union. They broadened my horizon and shared unforgettable times with me in my study abroad. In particular, I am grateful to Mr. Wenqi Niu, Mrs. Jing Zhao, Mr. Yongfeng Hou, Dr. Haodong Chang, Dr. Yu Suo, Dr. Yu Chen, Dr. Yan Wang, Dr. Kunpeng Zhao and Zhixun Zhao for their precious advice and encouragement in my Ph.D. research.

I gratefully acknowledge the funding from China Scholarship Council. Without its funding, my Ph.D work would not be possible.

Lastly, I would like to thank my family for all their selfless love and support. Without their encouragement, it would not be possible for me to conduct this research. In particular, my grateful thanks goes to my husband Dr. Chengzhang Zhu whose spiritual and academic support are so appreciated. Thank you.

Qian Li
University of Technology Sydney
July 2020

List of Publications

Related to the Thesis :

Published Papers :

1. **Li, Q.**, Wu, Q., Zhu, C., and Zhang, J., 2019, September. Bi-level masked multi-scale CNN-RNN networks for short text representation *In International Conference on Document Analysis and Recognition*. IEEE.
2. **Li, Q.**, Wu, Q., Zhu, C., Zhang, J. and Zhao, W., 2019, July. An inferable representation learning for fraud review detection with cold-start problem. *In 2019 International Joint Conference on Neural Networks*. IEEE.
3. **Li, Q.**, Wu, Q., Zhu, C., Zhang, J. and Zhao, W., 2019, April. Unsupervised user behavior representation for fraud review detection with cold-start problem. *In Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 222-236). Springer, Cham.
4. **Li, Q.**, Wu, Q., Liu, X., 2019. Multi-scale and hierarchical embedding for polarity shift sensitive sentiment classification. *Lecture Notes in Computer Science(including subseries Lecture Notes in Artifical Intellifence and Lecture Notes in Bioinformatics)*, 11633LNCS, pp.227-238.
5. Da, Q., Cheng, J., **Li, Q.**, Zhao, W., 2019. Socially-attentive representation learning for cold-start fraud review detection. *Communications in Computer and Information Science*, 1069, pp.76-91.

Submitted Manuscripts:

6. **Li, Q.**, Wu, Q., Zhu, C., Zhang, J., Zhao, W., 2020. Inferable representation learning for cold-start fraud review detection. *Neurocomputing* (Accepted).
7. **Li, Q.**, Wu, Q., Zhu, C., Cao, L., Zhang, J., 2020. Deep Non-IID Paragraph Representation for Sentiment Analysis. *IEEE Access* (Under Review).

8. Zhu, C., Cao, L., and **Li, Q.**,2020. Unsupervised hierarchical and heterogeneous coupling learning on dynamic categorical data. *Journal of Machine Learning Research* (Under Review).

OTHERS :

9. Zhao, W., **Li, Q.**, Zhu, C., Song, J., Liu, X. and Yin, J., 2018. Model-aware categorical data embedding: a data-driven approach. *Soft Computing*, 22, pp.3603-3619.
10. Zhu, C., Zhao, W.,**Li, Q.**, Li, P., Da, Q., 2019. Network embedding-based anomalous density searching for multi-group collaborative fraudsters detection in social media. *Computers, Materials and Continua*, 60(1), pp.317-333.
11. Xiang, L., Zhao, G., **Li, Q.**, Hao, W., Li, F., 2018. TUMK-ELM: A fast unsupervised heterogeneous data learning approach. *IEEE Access*, 6, pp.35305-35315.
12. Zheng, Y., Zhao, W., Sun, C., **Li, Q.**, 2019. Drug side-effect prediction using heterogeneous features and bipartite local models. *Computers, Materials and Continua*, 60(2), pp.481-496.
13. Cui, J., Long, J., Min, E., Liu, Q., **Li, Q.**, 2018. Comparative study of CNN and RNN for deep learning based intrusion detection system. *ICCCS 2018: Cloud Computing and Security (LNCS)*, 11067, pp. 159-170.

TABLE OF CONTENTS

List of Figures

List of Tables

Table **Page**