Identifying High-Value Social Entities from Twitter with Machine Learning and Multilingual Analysis

Siaw Ling Lo
BCM, MSc

A thesis submitted in total fulfilment

of the requirements for the degree of

Doctor of Philosophy

School of Electrical Engineering & Computing
Faculty of Engineering & Built Environment
The University of Newcastle
Callaghan, NSW 2308
Australia

March 2017

Statement of Originality

The thesis contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. I give consent to the final version of my thesis being made available worldwide when deposited in the University's Digital Repository, subject to the provisions of the Copyright Act 1968.

Signature: ……

Date: 26 March 2017

Acknowledgement of Collaboration

I hereby certify that the work embodied in this thesis has been done in collaboration with other researchers, or carried out in other institutions. I have included as part of the thesis a statement clearly outlining the extent of collaboration, with whom and under what auspices.

External supervisor: Assistant Professor Erik Cambria

Institution: Nanyang Technological University (NTU), Singapore

Collaboration: I collaborated with Prof Erik Cambria in the area of multilingual analysis research. A six-month attachment at NTU from July to December 2015 was completed to learn and exchange knowledge with the NTU research team under the guidance of Prof Erik Cambria. The following are the two journal papers published from the collaboration, with details documented in Chapters 5 and 6 of this thesis respectively:

- S. L. Lo, E. Cambria, R. Chiong, and D. Cornforth, "Multilingual sentiment analysis: from formal to informal and scarce resource languages," *Artif. Intell. Rev.*, pp. 1–29, 2016.
- S. L. Lo, E. Cambria, R. Chiong, and D. Cornforth, "A multilingual semi-supervised approach in deriving Singlish sentic patterns for polarity detection," *Knowl.-Based Syst.*, vol. 105, pp. 236–247, 2016

Signature: ……

Date: 26 March 2017

Acknowledgements

First, thanks to God. I can do nothing without His grace.

The work towards this thesis was financially supported by the International Postgraduate Research Scholarships and the Australian Postgraduate Awards. I would like to acknowledge the generous support of the Australian Government and The University of Newcastle, Australia, for covering my tuition fees and for providing me the research stipends. I consider myself blessed to have the opportunity to study at the Callaghan campus.

I would like to express my sincere gratitude to my main supervisor, Dr Raymond Chiong, for the continuous support of my PhD study both academically and emotionally. Specifically, I appreciate his patience and detailed guidance in writing (i.e., conference/journal papers and this thesis). He has imparted on me the essential skillset to write simple and yet have the ability to convey the message clearly. Without his guidance and constant feedback, this PhD would not be achievable.

My sincere appreciation also goes to my co-supervisor, Dr David Cornforth, for constantly encouraging me and providing me with many suggestions in improving my research. The enthusiasm he has shown in our discussions was contagious and it helped when problems were seen as learning opportunities. I would also like to acknowledge his advice on applying for an intellectual property disclosure for my approach in identifying high value social audience on Twitter.

I am thankful to have Dr Erik Cambria as my external supervisor and the opportunity to attach to his lab at Nanyang Technological University, Singapore, during my candidature. Besides learning the many facets of sentic computing, I also benefitted from the fruitful discussions with the team. It was one of the highlights in my PhD journey.

Besides my supervisors, I would like to thank the rest of my PhD committee: Dr Rukshan Athauda, Dr Suhuai Luo, Dr Peter Summons, and Dr Ilung Pranata, for their insightful comments and encouragements, and for the hard questions, which inspired me to widen my research from various perspectives.

I thank my fellow course mates for the stimulating discussions and for words of encouragements when things did not go as planned. In particular, I am grateful to Geoff for always being there to help. I will miss the jokes and humours we shared. I would also like to thank my friends from the Department of Chemistry and School of Education, Frem, Lek and Linda. This PhD journey is definitely a lot more meaningful and bearable because of the friendship and fun we have had. I am grateful to Dr Koh for helping to proofread some sections of this thesis and the many pieces of advice given during my candidature.

I am indebted to my parents for always believing in me and giving up many things for me to fulfil my dream to complete a PhD. In addition, I would like to say a heartfelt thank you to my in-laws and family members for helping in whatever way they could during my candidature. Em, you are my closest “family member” in Newcastle. Your support speaks volume when

things are challenging. I am grateful to have the prayer supports from my FOCUS friends, my life is richer because of all of you.

Finally, my deepest gratitude goes to my husband, Ling, for his unflagging love and faithful support in my life, especially during the candidature. And to my children, Xuan and Hur for loving and encouraging me with ways that melt my heart and kept me going.

Table of Contents

List of Figures

List of Tables

Abstract

With the vast amount, multilingual and real-time nature of social media data, it is challenging to extract relevant and useful information for individuals, companies and organisations. It is of interest to assess if the content shared and its multilingual expressions can be used to help a company in differentiating prospective customers from a general audience, or for individuals and organisations to detect and identify important topics that may otherwise go unnoticed within the mass of social media data. In this research, various methods and approaches have been investigated to identify high-value social entities in the form of social audiences and topics with minimal manual annotation effort. These include supervised machine learning methods such as the Support Vector Machine (SVM) ensemble, unsupervised clustering methods such as Latent Dirichlet Allocation (LDA), and text mining methods including latent semantic analysis and association rules. In addition, a hybrid framework has been developed for multilingual analysis by leveraging the strengths of both knowledge-based learning and machine learning. Twitter data, which is openly available, was used for validation and testing purposes.

Even though the aim of identifying high-value social audiences may seem to be different from that of identifying high-value topics, the underlying framework for the identification of these social entities remains the same. The first step is to earmark definitive contents that can provide information for constructing training or evaluation data with minimal annotation efforts. This step is crucial in order to avoid the alternative: the labour-intensive process of manually annotating data forming large online datasets. The second step is then to employ methods that are suitable to extract contents of interest. Both supervised and unsupervised methods such as the SVM ensemble and Twitter LDA have been used in this research to extract relevant social audiences. The SVM ensemble works well in this regard, as the contents of Twitter account owners are typically well-defined and can be used as training datasets for high-value target audience classification. On the other hand, since the number of classes or topics is not known, the unsupervised Dirichlet Process Mixture Model is instead preferred for topic detection. The third and last step is to assess the strengths and weaknesses of each method used in order to develop a hybrid approach. It is found that the combination or joint approach of various methods can often improve the recall and precision values and enable the identification of high-value social entities across datasets of different nature. This is supported by evidence from the promising results of a unique index devised for ranking high-value social audiences, which is called the *high-value social audience* (HVSA) index, on three different datasets, as well as the consistently higher precision and recall values from a 'Joint'

ranking method for identifying high-value topics with their sentiments in a huge set of multilingual tweets.

Methods and findings generated from this research have the potential to be adopted for addressing real-world problems. The HVSA index, for example, can be used to identify online customers who are highly likely to be interested in the content shared on social media by a business account owner. This can be useful in identifying prospective customers, or improving engagement with current customers. The ability to identify social media followers in a 'ranked' manner no doubt will help in better decision making, so that a (small) marketing budget can be spent more effectively. On the other hand, being able to detect high-value topics with their associated sentiments enables policy makers or organisations to understand issues of concerns on the ground and uncover possible actionable insights for a better community or customer reach.

List of Publications

A list of journal publications produced through this research:

1. S. L. Lo, R. Chiong and D. Cornforth, "Using Support Vector Machine Ensembles for target audience classification on Twitter", ***PLOS ONE*** 10(4):e0122855. DOI : 10.1371/journal.pone.0122855 (published 13 April 2015) http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0122855

2. S. L. Lo, R. Chiong and D. Cornforth, "Ranking of High-Value Social Audiences on Twitter", ***Decision Support Systems*** 85 (2016) 34-48. DOI : 10.1016/j.dss.2016.02.010 (published 2 March 2016).

3. S. L. Lo, E. Cambria, R. Chiong and D. Cornforth, "A Multilingual Semi-supervised approach in Deriving Singlish Sentic Patterns for Polarity Detection", ***Knowledge Based Systems*** 105 (2016) 236-247. DOI : 10.1016/j.knosys.2016.04.024 (published 26 April 2016)

4. S. L. Lo, E. Cambria, R. Chiong and D. Cornforth, "Multilingual Sentiment Analysis : From Formal to Informal and Scarce Resource Languages", ***Artificial Intelligence Review*** (2016) 1-29. DOI : 10.1007/s10462-016-9508-4 (published 20 August 2016)

5. S. L. Lo, R. Chiong and D. Cornforth, "An Unsupervised Multilingual Approach for Online Social Media Topic Identification", ***Expert Systems with Applications.*** DOI : 10.1016/j.eswa.2017.03.029 (published 21 March 2017)

A list of conference publications produced through this research:

1. S. L. Lo, D. Cornforth and R. Chiong, "Identifying the high-value social audience from Twitter through text-mining methods" in ***Proceedings of the 18th Asia Pacific Symposium on Intelligent and Evolutionary Systems*** Volume 1, 10-12 November 2014, pp. 325-339, Singapore. ISBN 978-3-319-13359-1_26

2. S. L. Lo, D. Cornforth and R. Chiong, "Effects of training datasets on both the Extreme Learning Machine and Support Vector Machine for target Audience Identification on Twitter" in ***Proceedings of ELM-2014*** Volume 1, 8-10 December 2014, pp.417-434, Singapore. ISBN 978-3-319-14062-9

3. S. L. Lo, D. Cornforth and R. Chiong, "Use of a high-value social audience index for target audience identification" in ***Proceedings of Artificial Life and Computational Intelligence***, 5-7 February 2015, pp.323-336, Newcastle, Australia. ISBN 978-3-319-14802-1

4. S. L. Lo, R. Chiong, D. Cornforth and Y. Bao "Topic Detection in Twitter via Multilingual Analysis" in ***Proceedings of Applied Informatics and Technology Innovation Conference***, 22-24 November 2016, pp. 1-22, Newcastle, Australia