

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

Statement of Originality.....	II
Acknowledgement of Collaboration	III
Acknowledgements.....	IV
Table of Contents.....	VI
List of Figures	XI
List of Tables	XIII
Abstract.....	XV
List of Publications	XVII
Chapter 1: Introduction	18
1.1 Background	18
1.2 Objectives.....	19
1.3 Research problems	19
1.4 Research contributions and publications	21
Chapter 2: Assessment of Machine Learning and Text Mining Methods for High-Value Social Audience Identification.....	24
2.1 Introduction	24
2.2 Related work.....	27
2.2.1 Customer segmentation and targeted marketing	27
2.2.2 HVSA identification on Twitter	28
2.3 The general combined approach	29
2.3.1 Seed words generation	30
2.3.2 Direct keyword match.....	31
2.3.3 Fuzzy keyword match.....	31
2.3.4 Twitter LDA	32
2.3.5 SVM	33
2.4 Machine learning approaches.....	35
2.4.1 ELM	35
2.4.2 SVM	36
2.5 Experimental setup	36
2.5.1 Data collection	36
2.5.2 Data cleaning and preparation	37
2.5.3 Performance metrics.....	38

2.5.4	Generation of training datasets	38
2.5.5	Generation of testing datasets	39
2.6	Results	40
2.6.1	Numbers of HVSA identified	40
2.6.2	Results of Twitter LDA.....	40
2.6.3	Results of SVM under the general combined approach	40
2.6.4	Comparison of various methods under the general combined approach.....	42
2.6.5	Training accuracy of various ELM configurations	42
2.6.6	Training accuracy of various SVM configurations.....	43
2.6.7	Comparing the ELM and SVM	44
2.6.8	Results of various methods.....	45
2.7	Discussion.....	46
2.8	Conclusion.....	49
Chapter 3: Identification of High-Value Social Audience through Ensemble Learning		50
3.1	Introduction	50
3.2	Related work	52
3.3	A new approach for HVSA classification	53
3.3.1	Discovery of followers' domains using Twitter LDA.....	53
3.3.2	SVM ensembles.....	57
3.4	Experimental setup	62
3.4.1	Performance metrics.....	62
3.5	Results.....	63
3.5.1	Representative target topical words identified	63
3.5.2	Results from individual SVMs of random sampling	64
3.5.3	Training performance of various SVM ensembles.....	65
3.5.4	Results of various SVM ensembles on the testing dataset.....	65
3.6	Discussion.....	66
3.7	Conclusion.....	68
Chapter 4: Ranking of High-Value Social Audience.....		69
4.1	Introduction	69
4.2	Related work	70
4.3	Details of datasets.....	71
4.3.1	Analysis of datasets.....	72
4.3.2	Dataset collection	72

4.3.3	Construction of testing datasets	73
4.4	HVSA identification	73
4.4.1	Fuzzy match.....	74
4.4.2	Twitter LDA	75
4.4.3	SVM ensembles.....	75
4.5	HVSA ranking.....	76
4.5.1	The count scoring schema	76
4.5.2	HVSA indices	78
4.6	Experimental setup and evaluation	79
4.6.1	Performance metrics.....	79
4.6.2	Ranking evaluation.....	79
4.7	Results.....	81
4.7.1	Thresholds derived for SVM ensembles with the count scoring schema	82
4.7.2	Performance of SVM ensembles on training datasets.....	83
4.7.3	Performance evaluation on the AF testing dataset	83
4.7.4	Ranking results based on the AF testing dataset	83
4.7.5	Results from the pooling strategy.....	89
4.8	Discussion.....	92
4.9	Conclusion.....	95
Chapter 5: Review of Multilingual Analysis with emphasis on Multilingual Sentiment Analysis		96
5.1	Introduction	96
5.2	Multilingual language processing	98
5.3	Current approaches used for multilingual sentiment analysis	101
5.3.1	Subjectivity analysis	102
5.3.2	Polarity analysis.....	103
5.4	Sentiment analysis on social media	115
5.4.1	English sentiment analysis on social media	116
5.4.2	Multilingual sentiment analysis on social media	117
5.4.3	Discussion.....	118
5.5	Work on scarce resource languages	118
5.5.1	Sentiment analysis	118
5.5.2	Speech recognition	120
5.5.3	Machine translation	120

5.6	Challenges and recommendations.....	120
5.6.1	Word sense dis-ambiguity	121
5.6.2	Language structure	121
5.6.3	Machine learning	121
5.6.4	Essential resources.....	122
5.6.5	A hybrid framework	123
5.6.6	Other considerations	124
5.7	Conclusion.....	125
Chapter 6: Deriving Singlish Sentic Patterns for Polarity Detection through a Semi-supervised Multilingual Approach		126
6.1	Introduction	126
6.2	Related work.....	128
6.3	Details of resources needed	129
6.3.1	Construction of a Singlish dictionary	129
6.3.2	Construction of the multilingual (English, Malay) and multifaceted polarity lexicon	130
6.3.3	Statistics of the Twitter dataset used	130
6.4	Methods and setups	131
6.4.1	Pre-processing of tweets	131
6.4.2	Construction of Singlish annotated testing datasets	132
6.4.3	Use of supervised machine learning	134
6.4.4	Construction of Singlish sentic patterns	135
6.4.5	The SinglishPD algorithm	139
6.4.6	A hybrid approach for polarity analysis using Singlish sentic patterns and machine learning.....	139
6.4.7	Performance evaluation.....	140
6.5	Results.....	141
6.5.1	Results of Singlish sentic patterns	141
6.5.2	Performance of various polarity assignment approaches based on Singlish annotated testing datasets	143
6.5.3	Results from the pooling strategy.....	144
6.6	Discussions and future plans	145
6.7	Conclusion.....	148
Chapter 7: Identification of High-Value Topics through an Unsupervised Multilingual Approach.....		150

7.1	Introduction	150
7.2	Related work	152
7.2.1	Topic detection	152
7.2.2	Multilingual analysis.....	156
7.3	Details of datasets.....	158
7.3.1	Twitter dataset collection	158
7.3.2	Singlish dataset construction	158
7.3.3	Ground truth dataset construction.....	159
7.4	Methods and experimental setups	159
7.4.1	Candidate day selection	160
7.4.2	Term ranking	161
7.4.3	Topic clustering	162
7.4.4	Evaluation.....	164
7.4.5	Multilingual sentiment analysis	166
7.5	Results	167
7.5.1	The list of matched terms and candidate days.....	167
7.5.2	Ground truth datasets analysis	171
7.5.3	Results of DPMM parameters optimisation	172
7.5.4	Evaluation via term recall and precision	173
7.5.5	Evaluation via topic recall and precision@10	176
7.5.6	Results of multilingual sentiment analysis.....	179
7.6	Discussion.....	181
7.7	Conclusion.....	183
	Chapter 8: Contributions and Future Research	185
8.1	Problem areas addressed.....	185
8.2	Thesis contributions.....	186
8.3	Directions for future research.....	187
	References	190

List of Figures

Figure 1. System architecture of the general combined approach	30
Figure 2. ROC curves based on testing data of various approaches using the SVM.....	42
Figure 3. ROC curves of various methods.....	43
Figure 4. F1 scores for the ELM and SVM based on top 10 score cut off.	44
Figure 5. F1 scores for the ELM and SVM based on top 30 score cut off.	45
Figure 6. ROC curves of various methods.....	46
Figure 7. Followers' domains discovery using Twitter LDA	54
Figure 8. The bootstrapping algorithm	57
Figure 9. A general architecture of bootstrapping using a single SVM model	58
Figure 10. A general architecture of the ensemble system using multiple SVM models.....	59
Figure 11. The majority vote algorithm	60
Figure 12. The bagging algorithm	61
Figure 13. The stacking algorithm.....	62
Figure 14. F measures of 10 SVM models generated from random samples.....	64
Figure 15. AUC of 10 SVM models generated from random samples.	64
Figure 16. ROC curves of various SVM ensembles on the testing dataset	66
Figure 17. Construction of various testing datasets	73
Figure 18. A simplified overall architecture for HVSA identification..	74
Figure 19. Construction of SVM ensembles.....	76
Figure 20. The threshold generation algorithm.....	78
Figure 21. Pooling strategy processes.....	81
Figure 22. ROC curves and sensitivity/specificity plots for the count scoring schema.....	84
Figure 23. ROC curves of various classifiers on a) samsungsg, b) ilovedealssg, and c) beaquafitness AF testing datasets.	85
Figure 24. AUC values of various classifiers for samsungsg, ilovedealssg and beaquafitness..	85
Figure 25. Evaluation results based on the metrics (P@k, AP@k, AP@all) using different AF testing datasets.....	86
Figure 26. Evaluation results based on the metrics (P@k, AP@k, AP@all) using the pooling strategy on different datasets.	86
Figure 27. Ranking performance of various methods based on the samsungsg AF testing dataset.	87
Figure 28. Ranking performance of various methods based on the ilovedealssg AF testing dataset.	87
Figure 29. Ranking performance of various methods based on the beaquafitness AF testing dataset.	88
Figure 30. Evaluation results based on the metrics (P@k, AP@k, AP@all) using the pooling strategy on different datasets.	89
Figure 31. Ranking performance of various methods based on ilovedealssg using the pooling strategy.	90
Figure 32. Ranking performance of various methods based on beaquafitness using the pooling strategy.	90
Figure 33. The recommended hybrid framework.....	124
Figure 34. Construction of Singlish annotated testing datasets.	132

Figure 35. The algorithm used to assign polarity to tweets.	133
Figure 36. The SVM model built with polarity emoticons.	135
Figure 37. The algorithm for extracting bigrams with polarity.	138
Figure 38. The SinglishPD algorithm.	139
Figure 39. Hybrid polarity analysis with Singlish sentic patterns and the SVM.	140
Figure 40. Overall architecture.	160
Figure 41. Candidate day selection.	160
Figure 42. Graphical model of DPMM.	164
Figure 43. Distribution of RTRate with respect to terms.	168
Figure 44. Distribution of InversedDiscussionRate with respect to terms and overlaying with RTRate.	169
Figure 45. Top three peaks identified in both datasets on term “riot” by Peak Identification ranking method.	171
Figure 46. The perplexity value generated by a range of alpha values.	173
Figure 47. The perplexity value and topic numbers generated by a range of beta value.	173
Figure 48. Results of term recall on Twitter LDA and K-Means clustering methods.	175
Figure 49. Results of term precision on Twitter LDA and K-Means clustering methods.	175
Figure 50. Results of term recall of the three clustering methods.	176
Figure 51. Results of term precision of the three clustering methods.	176
Figure 52. Results of topic recall of the three clustering methods.	178
Figure 53. Results of precision@10 of the three clustering methods.	178

List of Tables

Table 1. Types of training datasets and the number of features	39
Table 2. Types of testing datasets.....	39
Table 3. Numbers of HVSA identified by various methods.....	40
Table 4. Sample topic groups and their topical words	41
Table 5. SVM 10 fold cross-validation results	41
Table 6. Training accuracy of various ELM configurations.....	43
Table 7. Training accuracy of various SVM configurations	44
Table 8. AUC for various scoring schemas and training datasets	45
Table 9. AUC for various methods	46
Table 10. Interesting followers identified.....	47
Table 11. The domains identified from followers' tweets using Twitter LDA	56
Table 12. The configuration of bootstrapping using a single SVM model	58
Table 13. The configuration of various multiple SVM ensembles	59
Table 14. Topic groups identified via the seed words-fuzzy match approach and some of their topical words.....	63
Table 15. Results of 10 fold cross-validation of various SVM ensembles	65
Table 16. Results of various SVM ensembles on the testing dataset	66
Table 17. OpenCalais results for the three Twitter account owners (samsungsg, ilovedealssg and beaquafitness)	72
Table 18. Volume and period of datasets	73
Table 19. Threshold and AUC values for various SVM ensembles with the count scoring schema	82
Table 20. Results of 10 fold cross-validation for various SVM ensembles using the TF weighting schema	82
Table 21. Results of 10 fold cross-validation for various SVM ensembles using the TFIDF weighting schema	82
Table 22. Topic modelling based on the pooling strategy results.	91
Table 23. Audience segmentation using TLDA and the HVSA index on samsungsg followers .	92
Table 24. Tools for multilingual analysis.....	100
Table 25. Multilingual approaches used in subjectivity and polarity studies.....	109
Table 26. Lexicons and corpora used in multilingual sentiment analysis.....	113
Table 27. Singlish unigrams with polarity.....	141
Table 28. Singlish bigrams and trigrams with polarity.....	142
Table 29. Results based on the Singlish annotated testing datasets.....	143
Table 30. Results based on the pooled testing dataset with different types of sentic patterns .	144
Table 31. Classification results based on the pooled testing dataset.....	145
Table 32. Accuracy results from manual annotations.	146
Table 33. Results based on the emoticon dataset.....	147
Table 34. A comparison of topic detection approaches	154
Table 35. Top-10 candidate days with their matched terms and consolidated term frequency.	168

Table 36. Results of top 10 terms of each ranking method based on the Twitter dataset.	170
Table 37. OpenCalais results for the three candidate days.	172
Table 38. Results of topic recall and precision@10.	177
Table 39. Results of multilingual and English sentiment analysis.....	179
Table 40. Sample of random tweets detected by the multilingual polarity detection algorithm but not by the corresponding English-based polarity detection algorithm.....	180

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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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