University of Exeter
Department of Computer Science

Probabilistic topic models for sentiment analysis on the Web

Chenghua Lin

September 2011

Submitted by Chenghua Lin, to the the University of Exeter as a thesis for the degree of Doctor of Philosophy in Computer Science , September 2011.

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

(signature) ....

Dedicated to my parents, Xingwen Lin and Yuanfang Chen.

Acknowledgements

This thesis would not have been possible without the help and support from the following people.

First and foremost, I would like to thank my supervisor Yulan He, who has been an exemplary teacher and mentor in the past three years of my PhD studies. I am extremely grateful to Yulan's encouragement, support and patience during the difficult time of my research.

My gratitude also goes to my second supervisor Richard Everson, whose input and advice on my work were very useful. Richard is also very generous with his time for discussing research, for which I am grateful.

I also thank Alex Gong, without whose kind help I would not have been able to pursue the research I really enjoy.

I am fortunate to have a number of great friends who have supported me both research wise and non-research wise, particularly Naihui He, Junaid Khan, and Bing Qiao. Without them, my PhD would have been a less enjoyable one.

Deepest thanks to my parents and family. Their lifetime love and care help me regain my confidence and make me once again believe in myself. Finally, I thank Yuan. Her kindness, patience and care sustain me.

Abstract

Sentiment analysis aims to use automated tools to detect subjective information such as opinions, attitudes, and feelings expressed in text, and has received a rapid growth of interest in natural language processing in recent years. Probabilistic topic models, on the other hand, are capable of discovering hidden thematic structure in large archives of documents, and have been an active research area in the field of information retrieval. The work in this thesis focuses on developing topic models for automatic sentiment analysis of web data, by combining the ideas from both research domains.

One noticeable issue of most previous work in sentiment analysis is that the trained classifier is domain dependent, and the labelled corpora required for training could be difficult to acquire in real world applications. Another issue is that the dependencies between sentiment/subjectivity and topics are not taken into consideration. The main contribution of this thesis is therefore the introduction of three probabilistic topic models, which address the above concerns by modelling sentiment/subjectivity and topic simultaneously.

The first model is called the joint sentiment-topic (JST) model based on latent Dirichlet allocation (LDA), which detects sentiment and topic simultaneously from text. Unlike supervised approaches to sentiment classification which often fail to produce satisfactory performance when applied to new domains, the weakly-supervised nature of JST makes it highly portable to other domains, where the only supervision information required is a domain-independent sentiment lexicon. Apart from document-level sentiment classification results, JST can also extract sentiment-bearing topics automatically, which is a distinct feature compared to the existing sentiment analysis approaches.

The second model is a dynamic version of JST called the dynamic joint sentiment-topic (dJST) model. dJST respects the ordering of documents, and allows the analysis of topic and sentiment evolution of document archives that are collected over a long time span. By accounting for the historical dependencies of documents from the past epochs in the generative process, dJST gives a richer posterior topical structure than JST, and can better respond to the permutations of topic prominence. We also derive online inference procedures based on a stochastic EM algorithm for efficiently updating the model parameters.

The third model is called the subjectivity detection LDA (subjLDA) model for sentence-level subjectivity detection. Two sets of latent variables were introduced in subjLDA. One is the subjectivity label for each sentence; another is the sentiment label for each word token. By viewing the subjectivity detection problem as weakly-supervised generative model learning, subjLDA significantly outperforms the baseline and is comparable to the supervised approach which relies on much larger amounts of data for training.

These models have been evaluated on real world datasets, demonstrating that joint sentiment topic modelling is indeed an important and useful research area with much to offer in the way of good results.

Publications

Chapter 3 is based on the previously published work:

Lin, C., He, Y., Everson, R. and Rüger, S. Weakly-supervised Joint Sentiment-Topic Detection from Text, *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, to appear.

Lin, C., He, Y., and Everson, R. A Comparative Study of Bayesian Models for Unsupervised Sentiment Detection, In *Proceedings of the 14th Conference on Computational Natural Language Learning (CoNLL)*, Uppsala, Sweden, 2010.

Lin, C. and He, Y. Joint Sentiment/Topic Model for Sentiment Analysis, In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM)*, Hong Kong, China, 2009.

Chapter 5 is based on the previously published work:

Lin, C., He, Y. and Everson, R. Sentence Subjectivity Detection with Weakly-Supervised Learning, In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP)*, Chiang Mai, Thailand, 2011.

Other publications:

He, Y., Lin, C. and Alani, H. Automatically Extracting Polarity-Bearing Topics for Cross-Domain Sentiment Classification, In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, Portland, Oregon, 2011.

He, Y. and Lin, C. Protein-Protein Interactions Classification from Text via Local Learning with Class Priors, In *Proceedings of the 14th International Conference on Applications of Natural Language to Information Systems (NLDB)*, Saabrucken, Germany, 2009.

Contents

List of Figures

List of Tables