Abstract

Microblogging services have been becoming increasingly popular over the last decade. Many people express their opinions and feeling about anything in the famous microblogging service, Twitter. These people's opinions can be grabbed easily and publicly through the interface provided by Twitter. Both individuals and organization are increasingly using this data for decision making. Customers want to know the opinion of other users before making purchase decision. Companies want to know the feedback from users about their products and also their competitors. Therefore, opinion mining and sentiment analysis become one of the major research topics in the field of natural language processing. Early work on the sentiment analysis proposed methods of classifying the sentiment on the traditional social network, i.e. forum, webboard and review. However, it is more difficult to analyze sentiments on tweets. Tweets are very short and contain a lot of informal expressions, i.e. slang, emoticon, typographical errors and a lot of words that are not compiled in a dictionary. The solution and method of traditional sentiment analysis system cannot be applied directly because of these unique characteristic of the microblogging. Moreover, existing sentiment analysis approaches mainly focus on measuring the sentiment of individual tweets or predict the massive opinions for a specific target. However, identification of the opinions of individual users is another important task that is often required in practical opinion mining systems.

In this thesis, we research and develop several methods of classifying the sentiments on microblogging, aiming to address the problems described above. We do not only focus on classifying the sentiment of each tweet by considering the textual information, which is usually short and hard to interpret. We aim to seek other characteristics in microblogging to extract the extra knowledge for boosting the performance of the sentiment analysis. Three main sentiment analysis tasks are considered, namely tweet-level sentiment analysis, target-dependent sentiment analysis, and user-level sentiment analysis. In the following, we describe each analysis one-by-one.

First, in the tweet-level sentiment analysis, we introduce a hybrid approach that uses a

lexicon for sentiment words to alleviate the data sparseness problem inherent in machine learning approaches and improve sentiment classification in tweets. The data sparseness problem can be reduced by the following two methods. We first estimate the potential polarity of objective and out-of-vocabulary (OOV) words and use these words as additional information of the existing sentiment lexicon. The polarity scores of OOV words are estimated based on the assumption that the polarities of words are coincident with the polarity of their associated sentences, using a collection of labeled sentences with their polarity. Then, we introduce a novel feature weighting method by interpolating sentiment lexicon score into uni-gram score in the feature vectors of SVM.

Second, in the target-dependent sentiment analysis, we propose a method for incorporating on-target sentiment information and user sentiment information into a machine learning classifier for the target-dependent sentiment analysis of the tweets. Three extra resources, the add-on lexicon, the extended target list, and the competitors list, are automatically constructed from the unlabeled tweets. The target specific training data is created based on heuristic rules and the lexicon-based sentiment analysis method. Two new features for training the sentiment classifier are introduced. One is the on-target sentiment feature, giving greater weight to the sentiments of the words near the target; the other is the user sentiment feature that captures the tendency of the sentiment expressed by the same user.

Third, in the user-level sentiment analysis, we propose a novel graph-based method that incorporates the information of both textual information, as well as the explicit and implicit relationships between the users, into a heterogeneous factor graph for the sentiment analysis of the tweets at the user level. Our framework takes into consideration not only the explicit connections such as follow, mention and retweet but also the implicit connections between users. An implicit connection refers to the relations of users who share similar topics of interest. The implicit relations among the users are extracted from their historical tweet corpus. Since the presence of the explicit relations in some social network is limited, the implicit relations allow us to utilize the data in social network more effectively. We also propose a new enhanced pooling method, "Hashtag-PMI", to more precisely infer the latent topics by the conventional LDA (Latent Dirichlet Allocation) from the tweet corpus.

Both public and real-life tweet corpora are used in our experiments. The results of experiments show that our method achieves 64-70%, 59-62% and around 65% accuracy on the tweet-level, target dependent and user-level sentiment analysis tasks respectively. The proposed method is effective and significantly improves the performance compared to the several baselines and existing methods.

Keywords: Sentiment Analysis, Machine Learning, Knowledge Acquisition, Topic Modeling, Microblogging