Neurocomputing 175 (2016) 599–609

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Structured microblog sentiment classification via social context regularization

Fangzhao Wu [a,*], Yongfeng Huang [a], Yangqiu Song [b]

[a] Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
[b] Lane Department of Computer Science and Electrical Engineering, West Virginia University, USA

ARTICLE INFO

Article history:
Received 19 June 2015
Received in revised form
7 September 2015
Accepted 28 October 2015
Communicated by Y. Chang
Available online 3 November 2015

Keywords:
Sentiment classification
Microblog
Social context

ABSTRACT

Microblog sentiment analysis is a fundamental problem for many interesting applications. Existing microblog sentiment classification methods judge the sentiment polarity mainly according to textual content. However, since microblog messages are very short and noisy, and their sentiment polarities are often ambiguous and context-dependent, the accuracy of microblog sentiment classification is usually unsatisfactory. Fortunately, microblog messages lie in social media and contain rich social contexts. The social context information often implies sentiment connections between microblog messages. For example, a microblogging user usually expresses the same sentiment when posting multiple messages towards the same topic. Motivated by these observations, in this paper we propose a structured microblog sentiment classification (SMSC) framework. Our framework can combine social context information with textual content information to improve microblog sentiment classification accuracy. Two kinds of social contexts are used in our framework, i.e., social connections between microblog messages brought by the same author and social connections brought by social relations between users. In our framework, social context information is formulated as the graph structure over the sentiments of microblog messages. The objective function of our framework is a tradeoff between the agreement with content-based sentiment predictions and the consistency with social contexts. An efficient optimization algorithm is introduced to solve our framework. Experimental results on two Twitter sentiment analysis benchmark datasets indicate that our method can outperform baseline methods consistently and significantly.

1. Introduction

With the development of Web 2.0, microblogging websites, such as Twitter[1] and Weibo,[2] have gained huge popularity and attracted hundreds of millions of users. People post massive short messages everyday to release latest news and share their opinions on various topics. Analyzing the sentiments in these large-scale microblog messages can help sense public's opinions on products, companies, political events and so on, which has wide applications [1]. For example, if we find a user who shows special favor to a certain brand, e.g., Apple, by analyzing his/her tweets, we can make more useful recommendations or more effective advertisements for him/her. Mining customers' opinions towards a newly released product from their microblogs can help improve product quality, design more successful advertising campaign, and conduct more timely crisis management [1,2]. Analyzing massive voters' opinions from their microblogs is also important for improving a political campaign [3,4]. Thus microblog sentiment classification has become a hot research topic in both academic and industrial fields in recent years [1,2,5–7].

Existing microblog sentiment classification methods judge the sentiment of a message mainly according to the features extracted from the textual content [8–10]. However, microblog messages are very short and usually consist of only one or two sentences, even several words [6]. Thus it is quite difficult to gain sufficient statistical information from the textual content to classify the sentiment accurately. Moreover, microblog messages' sentiment polarities are often ambiguous and context-dependent. The same message may show different sentiments in different contexts [11]. For example, a tweet may be "@Obama you really care about U.S. people!" If this tweet is in the context of health care reform, it may convey positive sentiment. However, if in the context of PRISM (NSA's surveillance project), it is probably ironic and conveys

* Corresponding author.
E-mail address: wfz12@mails.tsinghua.edu.cn (F. Wu).
[1] https://www.twitter.com/
[2] http://www.weibo.com/

http://dx.doi.org/10.1016/j.neucom.2015.10.101
0925-2312/

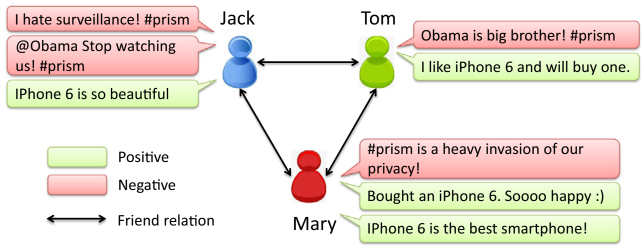

Fig. 1. A toy example of people's opinions towards "PRISM" and "iPhone 6" topics in social media. The microblog messages posted by the same user or by friends tend to convey similar sentiments towards the same topic.

negative sentiment. Thus, it is quite challenging to classify the sentiments of microblog messages purely using the textual content and the classification accuracy is usually not good enough for real applications [6].

Fortunately, microblog messages lie in the environment of social media and contain rich social context information. These social contexts contain useful clues of sentiment connections between microblog messages [6,7,1]. For example, a person's opinion towards the same target/topic usually keeps consistent within a short period. This has been verified by social science theories such as *sentiment consistency* [12]. In addition, the opinions of friends towards the same topic usually tend to be similar. This phenomenon also has been validated in psychological fields and formulated as the principle of *homophily* [13] or "birds of a feather flock together" [14]. Therefore, the microblog messages posted by the same user or by users with friend relations tend to convey similar sentiments towards the same topic. Fig. 1 shows an illustrative example. Thus social contexts can provide additional information for microblog sentiment classification.

One way to use social contexts in microblog sentiment classification is encoding them into sentiment classification model learning [15,6]. For example, Hu et al. proposed to incorporate social contexts into the process of learning a Least Square sentiment classifier [6]. In their method, the sentiment classification model is constrained to assign similar sentiment scores to tweets with social connections at the model learning stage. However, these methods are still content-based sentiment classification methods, because when classifying the sentiments of unseen microblogs, only textual content is used.

In this paper, we propose a new way to exploit social context information for microblog sentiment classification. Specifically, we propose a structured microblog sentiment classification (SMSC) framework which can combine social context information with textual content information to classify the sentiments of microblog messages more accurately. In our method, the social contexts are used at the prediction stage and formulated as the graph structure over the sentiments of the microblog messages. Two kinds of social contexts, i.e., social context introduced by the same author (denoted as *user context*), and social context introduced by users with friend relations (denoted as *friend context*), are used in our framework. The objective function of our framework is the tradeoff between consistency with content-based sentiment predictions and agreement with social context information. We propose an efficient algorithm based on alternating direction method of multipliers (ADMM) [16] to solve the optimization problem in our framework and introduce an accelerated method based on fast iterative shrinkage thresholding algorithm (FISTA) [17] to tackle the most time-consuming step in our algorithm. Experimental results on two Twitter sentiment analysis benchmark datasets validate the effectiveness and efficiency of our method.

The remainder of this paper is organized as follows. In Section 2, we introduce several related works. In Section 3, we discuss how to extract and formulate the social contexts for microblog sentiment classification. In Section 4, we introduce our structured microblog sentiment classification framework and the optimization algorithms in detail. In Section 5, we present the experimental results. In Section 6, we conclude this paper and outline several future works.

2. Related work

Sentiment analysis has been studied for several years and gained success in analyzing the sentiment of reviews [18–20], blogs [21,22], news articles [23] and so on. Sentiment classification methods can be divided into two major categories, i.e., lexicon based methods and machine learning based methods [24,25]. Lexicon based methods utilize a general or domain-specific sentiment lexicon to identify positive and negative sentiment words in a document [26]. Then the overall sentiment polarity of the document is judged by summarizing the sentiment polarities of these sentiment words. In machine learning based methods, the sentiment classification problem is regarded as a text classification problem [18,8,27]. A sentiment classification model is first trained using a labeled sentiment dataset. Then this model is used to assign sentiment labels to unseen documents. Both lexicon based methods and machine learning based methods are purely content-based sentiment classification methods.

Microblog sentiment analysis is a hot research topic in these years [8,28,6,10]. Since microblog messages are short and noisy, containing massive informal words, microblog sentiment classification is more challenging than classifying the sentiments of long texts [6]. In order to reduce the costly and time-consuming manual labeling work, Go et al. proposed to use emoticons as noisy sentiment labels to automatically train sentiment classifiers [8]. In order to identify the sentiment polarities of massive informal words in microblog messages, Kiritchenko et al. proposed to build Twitter-specific sentiment lexicons using associations between words and emoticons [10]. Then sentiment features, for example, the numbers of positive and negative sentiment words, were extracted for each microblog message according to these sentiment lexicons. These sentiment features are combined with textual content features, such as n-gram and POS tag features, to train a sentiment classifier using a labeled dataset [10]. Although these methods tackle some of the difficulties in microblog sentiment classification, they are still purely content-based methods. The fact that microblog messages are networked data is not considered in these methods. As discussed in previous section, due to the characteristics of microblogs, there are still many cases that content-based methods cannot handle very well.

More recently, researchers have recognized the importance of social context information in sentiment analysis and opinion mining tasks [29,30,15,6,7,1]. For example, Ren and Wu tried to incorporate the friend relations between users and content similarities between topics into the matrix factorization framework of a recommendation system to predict an individual's opinion towards an unseen but related topic [1]. Similarly, Tan et al. proposed to use follow relations and mention relations (extracted using "@") to help infer Twitter users' sentiments towards different topics [7]. These methods are user-level or user-topic level sentiment classification approaches, while our method is message-level and more fine-grained. In [6], Hu et al. proposed to use social contexts as additional information for supervised sentiment classification model learning. In their method, the sentiment classification model is constrained to assign similar sentiment scores to microblog messages connected by social contexts at the model learning stage. However, their method is still a purely content-based approach, because when judging the sentiment polarity of an unseen microblog message, only the textual content is used.