Decision Support Systems 87 (2016) 39–49

Contents lists available at ScienceDirect

Decision Support Systems

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

Towards building a high-quality microblog-specific Chinese sentiment lexicon

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

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

ARTICLE INFO

Article history:
Received 18 December 2015
Received in revised form 8 April 2016
Accepted 27 April 2016
Available online 4 May 2016

Keywords:
Sentiment lexicon
Sentiment analysis
Microblog

ABSTRACT

Due to the huge popularity of microblogging services, microblogs have become important sources of customer opinions. Sentiment analysis systems can provide useful knowledge to decision support systems and decision makers by aggregating and summarizing the opinions in massive microblogs automatically. The most important component of sentiment analysis systems is sentiment lexicon. However, the performance of traditional sentiment lexicons on microblog sentiment analysis is far from satisfactory, especially for Chinese. In this paper, we propose a data-driven approach to build a high-quality microblog-specific sentiment lexicon for Chinese microblog sentiment analysis system. The core of our method is a unified framework that incorporates three kinds of sentiment knowledge for sentiment lexicon construction, i.e., the word-sentiment knowledge extracted from microblogs with emoticons, the sentiment similarity knowledge extracted from words' associations among all the messages, and the prior sentiment knowledge extracted from existing sentiment lexicons. In addition, in order to improve the coverage of our sentiment lexicon, we propose an effective method to detect popular new words in microblogs, which considers not only words' distributions over texts, but also their distributions over users.The detected new words with strong sentiment are incorporated in our sentiment lexicon.We built a microblog-specific Chinese sentiment lexicon on a large microblog dataset with more than 17 million messages. Experimental results on two microblog sentiment datasets show that our microblog-specific sentiment lexicon can significantly improve the performance of microblog sentiment analysis.

1. Introduction

Microblogging services, such as Twitter[1] and Weibo[2] , have become increasingly popular in recent years.They provide great platforms to hundreds of millions of users to freely express their opinions on various topics, such products, brands and companies, in an unconstrained and unbiased environment [1]. Thus microblogging platforms have become ideal sources of customer opinions and market intelligence [2]. Analyzing and summarizing the sentiments in these large-scale opinion-rich microblogs can provide useful knowledge to both companies and customers for making better decisions [2,3,4,5]. For example, consumers can make more informed decisions when buying products or services by referring to masses of other customers' opinions and perform comparison shopping [6]. Companies can sense how and in which aspects their customers like and dislike their products or services in real time. They can remain competitive by inferring customers' need and taste from their microblogs, and manage good relationships with customers by analyzing and responding to their comments timely [2]. In addition, companies can improve their advertising campaigns and market strategies by analyzing their customers' opinions towards their brands and products on a large scale [3]. Besides, companies managers can track the fluctuations of their companies' reputation as well as those of their competitors' to adjust their producing and sale strategies [7]. However, since microblogs are on an extremely large scale, it is costly and time-consuming to analyze the opinions in them by manual inspection. Thus, sentiment analysis systems which can aggregate, organize, analyze and summarize the opinions in microblogs automatically is very important for extracting useful knowledge from massive opinion-rich microblogs in real time and on a large scale, in order to help decision support systems and decision makers to make better decisions in business activities [2].

* Corresponding author.
E-mail address: wfz12@mails.tsinghua.edu.cn (F. Wu).

[1] https://twitter.com/.
[2] http://www.weibo.com/.

http://dx.doi.org/10.1016/j.dss.2016.04.007
0167-9236/

Sentiment lexicon, which consists of a list of sentiment words and phrases as well as their sentiment polarities and intensities, is the most important component in sentiment analysis systems [3,8–11]. The performance of sentiment analysis systems heavily depends on the accuracy and coverage of the sentiment lexicon they use. However, existing sentiment lexicons are not suitable for microblog sentiment analysis due to two reasons. First, when posting microblog messages, users frequently use informal new words, such as "tnx" and "coooool", to express their emotions. Many of these informal new words convey rich sentiment information and are important for microblog sentiment analysis. But they are not covered by traditional sentiment lexicons [12]. Second, formal words may have different sentiments in microblogging scenario. For example, Chinese word "鸭梨" represents a kind of pear in traditional texts. However, it is often used to express "pressure" in microblog messages. Manually detecting and annotating these informal new words and formal words with changed sentiments are costly and time-consuming, because they are on a large scale and continuously emerging. Thus an automatic method to build a microblog-specific sentiment lexicon is of great value [12].

Several methods have been proposed for English microblog-specific sentiment lexicon construction. For example, Kiritchenko et al. proposed to calculate words' sentiment scores by leveraging their associations with emoticons (such as ":)") or sentiment-word hashtags (such as #joy) [13]. Tang et al. regarded the sentiment lexicon construction as a word-level sentiment classification problem, and proposed a representation learning method to classify sentiments of English n-grams in tweets [12]. Compared to English lexicons, building microblog-specific Chinese sentiment lexicon is more challenging because there is no natural segmentation symbol such as blank space to separate continuous Chinese characters into words. Existing Chinese word segmentation tools often fail to detect the user-invented new words used in microblogs and simply split them into single characters [14]. Thus, the aforementioned methods designed for English sentiment lexicon construction cannot be directly applied in our task.

To build a microblog-specific Chinese sentiment lexicon, Feng et al. inferred words' sentiment scores using their associations with positive and negative sentiment emoticons [15]. However, this method can neither detect new words used in Chinese microblogs nor specify sentiment polarities for them. Thus the coverage of sentiment lexicon built using this method is still limited. More recently, Huang et al. proposed a pattern-based method to detect adjective new words from POS-tagged microblog texts. They computed the sentiment polarities of these words using their associations with sentiment emoticons [14]. They found that incorporating these new sentiment words into traditional sentiment lexicons can benefit Chinese microblog sentiment classification. However, this method still has several limitations. First, it can only detect adjective new sentiment words while many new sentiment words belong to other syntactic classes, such as verb and noun, which limits the coverage of the sentiment lexicon built in this way. Second, since microblog texts are very casual and noisy, it is difficult to obtain high-quality POS-tagging results. And finally, traditional words may change their sentiments when used in microblogging scenario.

To overcome these limitations, in this paper we propose a data-driven approach to build a high-quality microblog-specific Chinese sentiment lexicon.

First, in order to improve the coverage of our sentiment lexicon, we develop an effective new word detection method to detect the popular user-invented new words used in microblogs. The major feature of this method is that it utilizes not only words' distributions over messages but also their distributions over users. It iteratively detects candidate new words and adds them to the dictionary of Chinese word segmentation tools. The candidate new words detected in previous iterations are refined in subsequent iterations. Accordingly, the word segmentation performance can be improved and more new words can be detected simultaneously.

Second, we propose a unified framework to build a high-quality microblog-specific Chinese sentiment lexicon by incorporating three kinds of sentiment knowledge. The first one is the word-sentiment knowledge, which represents words' sentiment scores extracted from the associations between words (both formal and new words) and emoticons. The second one is the sentiment similarity knowledge, which represents the sentiment similarities among words, and is extracted from all the available microblog messages. The third one is the prior knowledge extracted from existing sentiment lexicons.

We build a microblog-specific Chinese sentiment lexicon using a large Chinese microblog dataset with more than 17 million messages. Experimental results on two Chinese microblog sentiment datasets validate that our microblog-specific sentiment lexicon can outperform existing sentiment lexicons by a large margin in various sentiment analysis tasks, such as subjectivity detection and sentiment polarity classification, at both sentence and message levels.

The major contributions of our work are as follows:

- We propose an effective and purely data-driven method to detect popular user-invented new words in Chinese microblogs, which utilizes both words' distributions over text messages and their distributions over users. These new words can improve the coverage of our microblog-specific sentiment lexicon significantly.
- We propose a unified framework which can incorporate three kinds of sentiment knowledge to build a high-quality microblog-specific sentiment lexicon.
- We build a microblog-specific Chinese sentiment lexicon using a large microblog dataset, and conduct extensive experiments on two microblog sentiment datasets to evaluate its performance in various sentiment analysis tasks.

The rest of this paper is organized as follows. Related works are introduced in Section 2. In Section 3, we introduce our new word detection method. We present our approach to microblog-specific sentiment lexicon construction in Section 4. We report the experimental results in Section 5. Section 6 concludes this paper.

2. Related work

In this section we introduce several works related to sentiment lexicon construction and Chinese new word detection.

2.1. Sentiment lexicon construction

Sentiment lexicons, which consist of a list of sentiment words as well as their sentiment polarities and intensities, play an important role in many sentiment analysis systems [9,10]. Traditional sentiment lexicons were constructed manually [16] or automatically [12,13,17–21]. In automatic methods, usually a set of seed sentiment words and their sentiment labels are given in advance. Then the information in these seed sentiment words is propagated to other words. Many methods following this research direction are based on graph propagation [18,19]. In these methods, a sentiment similarity graph is first constructed, where words and phrases are modeled as nodes and sentiment connections between them are regarded as edges. These sentiment connections can be extracted from thesauruses [8,18], syntactic contexts [19], parsing results [22], and so on. For example, Hu and Liu exploited the synonym and antonym relations in WordNet to construct the sentiment similarity graph [18]. Esuli and Sebastiani used the glosses in WordNet to infer the words' sentiment relations [8]. Wiebe utilized the dependency