Knowledge-Based Systems 000 (2016) 1–9

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Word sense disambiguation based sentiment lexicons for sentiment classification

Chihli Hung*, Shiuan-Jeng Chen

Department of Information Management, Chung Yuan Christian University, Taiwan

ARTICLE INFO

Article history:
Received 18 April 2016
Revised 19 June 2016
Accepted 23 July 2016
Available online xxx

Keywords:
Sentiment analysis
Opinion mining
Word sense disambiguation
SentiWordNet
Sentiment lexicon

ABSTRACT

Sentiment analysis has attracted much attention from both researchers and practitioners as word-of-mouth (WOM) has a significant influence on consumer behavior. One core task of sentiment analysis is the discovery of sentimental words. This can be done efficiently when an accurate and large-scale sentiment lexicon is used. SentiWordNet is one such lexicon which defines each synonym set within WordNet with sentiment scores and orientation. As human language is ambiguous, an exact sense for a word in SentiWordNet needs to be justified according to the context in which the word occurs. However, most sentiment-based classification tasks extract sentimental words from SentiWordNet without dealing with word sense disambiguation (WSD), but directly adopt the sentiment score of the first sense or average sense. This paper proposes three WSD techniques based on the context of WOM documents to build WSD-based SentiWordNet lexicons. The experiments demonstrate that an improvement is achieved when the proposed WSD-based SentiWordNet is used.

1. Introduction

This paper investigates the disambiguation of ambiguous words and builds domain oriented sentiment lexicons based on a well-known sentiment lexicon, the SentiWordNet [3], for the task of word-of-mouth (WOM) sentiment classification. With the development of the internet, online forums, micro-blogs, blogs, social networks and web platforms have become a primary channel for users to share their personal experiences, feelings and opinions regarding products, services, brands and events with family, friends, and the general public. This information includes large amounts of product WOM documents, which have long been a major influence on decision-making in consumer purchasing behaviors [6]. Through the sharing of WOM documents, consumers are able to reference the experiences and recommendations of others to help them decide whether or not they want to purchase a product or service [27]. For businesses and commercial entities, these WOM documents provide a crucial means to understanding the views and preferences of consumers [35]. In contrast to traditional market surveys, WOM document analysis can provide information that is both closer to real-time and to consumers' opinions, and can thus create new opportunities and a competitive advantage [8,32,65].

The primary goal of sentiment analysis is to employ automated methods to extract positive, negative, or neutral emotions from WOM documents, while determining the overall feeling that the writer of the document wishes to express, thereby extracting implicit information from the text. In the literature, sentiment analysis frequently utilizes a sentiment lexicon to help identify the words used in documents that reflect sentiment. The SentiWordNet sentiment lexicon uses WordNet [42] as its foundation and takes a semi-supervised approach toward constructing a vocabulary database that includes the ability to determine the emotional polarity of words.

Although SentiWordNet can help identify sentimental words used in WOM documents, it can only be considered a general sentimental vocabulary database. This is due to the fact that when using SentiWordNet to conduct sentiment analysis, its accuracy is still affected by the issue of word sense identification. Many words have more than one meaning and the meanings expressed by words will change based on different backgrounds and environments [43]. When the meanings of words change, the sentimental attitudes that they express also change. Therefore the ability to select the correct meaning represented by the word in a particular text will also affect the effectiveness of the sentimental analysis results.

In SentiWordNet each meaning of a word is considered a sense, and each sense is given a corresponding sentiment polarity score. Sense 1 represents the sense used most often in general situations [1,24,38,43]. There are two common approaches to the selection of sentiment scores for words that contain multiple meanings. The first method is to directly pick sense 1 of the word to serve as the

* Corresponding author. Fax: 88632655499.
E-mail address: chihli@cycu.edu.tw (C. Hung).

http://dx.doi.org/10.1016/j.knosys.2016.07.030
0950-7051/© 2016 Elsevier B.V. All rights reserved.

meaning of the word in the text (e.g. [29]). This method does not consider the impact of domain knowledge and may result in biased sentiment analysis results. For example, in the context of movie reviews, the word “suck” is most likely to mean “inappropriate or lousy”, and is used to express the opinion that the movie is very poor. The sentiment, in this case, is negative. However, in SentiWordNet, the meaning represented by sense 1 of the word is “a sucking action” and is classified as having neutral sentiment. If the sense 1 meaning is automatically selected, the sentiment would clearly be incorrect. The second method is to take the average of the sense scores for all meanings for words with multiple meanings, and use the average sentiment score to conduct analysis (e.g. [44]). However, this method does not take the effects of domain knowledge into consideration, and could also result in issues with the accuracy of sentiment analysis.

Therefore this paper proposes an approach in which word sense disambiguation (WSD) techniques are applied to movie and hotel review documents, to revise SentiWordNet 3.0 [3] from a general purpose sentiment lexicon into a movie domain oriented sentiment lexicon and a hotel domain oriented sentiment lexicon, in order to improve the performance of sentiment classification for the domain documents.

The remainder of this paper is organized as follows. In Section 2, we briefly review related work including sentiment analysis and word sense disambiguation. Section 3 introduces the approach to building a WSD-based sentiment lexicon. Section 4 shows the experiment design and results. Finally, a conclusion and possible future work are presented in Section 5.

2. Related work

Sentiment analysis is divided into the following six tasks: sentiment classification, subjectivity classification, opinion summarization, opinion retrieval, sarcasm and irony, and others [54]. Most sentiment analysis tasks focus on the sentiment classification of WOM documents according to their sentiment polarity, i.e. positive or negative [18,19,26,29,34,50,54,58].

These tasks can generally be completed more efficiently when an accurate and large-scale sentiment lexicon is used [22,30,54]. Various researchers have developed a number of sentiment lexicons, such as General Inquirer (GI) [57], WordNet-Affect [59], SentiWordNet [3,23], SenticNet [9,10,12]. Of these lexicons, SentiWordNet, based on the online English dictionary, WordNet [42], has become a frequently used sentiment lexicon due to its large-scale coverage. SentiWordNet defines each synonym set (or synset) in WordNet with three sentiment labels: positivity, neutrality and negativity. Each label has a specific value between zero and one, and the sum of the three labels is equal to one. Many sentiment analysis models are developed based on SentiWordNet [20,29,44,45,52,54]. For example, Ohana and Tierney [45] evaluated the function of sentimental scores in SentiWordNet for the automatic sentiment classification of film reviews. Their proposed approach using the sentiment values in SentiWordNet performs slightly better than the approach in which only the frequency of sentimental words is used. Saggion and Funk [52] applied SentiWordNet to opinion classification for a business-related data source. Devitt and Ahmad [21] classified financial and economic news based on SentiWordNet and analyzed whether or not these sentimental documents influenced the market. Hung et al. [30] applied SentiWordNet for tagging sentimental orientations and classifying documents into five qualitative categories. Hung and Lin [29] revised the sentiment scores for neutral words defined in SentiWordNet and obtained an improved sentiment classification performance. More complete literature reviews in the field of sentiment analysis or opinion mining can be found in [4,11,40,48,54,58].

One significant step in sentiment analysis is the discovery of sentimental words [64]. As many words in SentiWordNet contain more than one sense, they must be interpreted according to the context in which they occur. The task of automatic identification of meaning for words in context is called word sense disambiguation (WSD) [43]. WSD is an historical task in the field of natural language processing and has been used in various applications such as spam filtering, document classification, information retrieval, etc. [2,15,25,31,37,41,43,49,55,60]. For example, the traditional Lesk algorithm [39] disambiguates words in short phrases, based on the greatest number of common words shown in the definition sentence of each word in the same phrase. This algorithm looks up traditional dictionaries, such as the Oxford Advanced Learner’s Dictionary, for word definition. Banerjee and Pedersen [5] modified the traditional Lesk algorithm and proposed the adapted Lesk algorithm. The adapted Lesk algorithm disambiguates words in sentences. It compares the target word and its surrounding words by the glosses of their synonymous sets defined in WordNet.

However, most sentiment-based classification tasks extract sentimental words from SentiWordNet without dealing with word sense disambiguation, but directly adopt the sentiment score of the first sense or average sense [29,44,54]. Unlike existing work in the literature, this paper focuses on the issue of ambiguous words and proposes three WSD techniques for improvement of the performance of sentiment classification by re-ranking the sentiment order in SentiWordNet.

3. Methodology

This paper recognizes that words used in different domains may have different senses, different sentiment values and even different sentiment orientations. We propose three methods to revise a general sentiment lexicon, SentiWordNet, into a WSD-based or domain oriented sentiment lexicon, in order to improve its effectiveness for sentiment classification. Fig. 1 shows the structure of the proposed approaches. Our methodology is divided into four phases, which are preprocessing of WOM documents, tokenization, word sense disambiguation and building the WSD-based sentiment lexicon. Finally, the model that extracts sentiments from the traditional SentiWordNet for the unseen test set is treated as the benchmark. We then compare the benchmark with the proposed model for the unseen test set that extracts sentiments from the WSD-based SentiWordNet.

3.1. Preprocessing of WOM documents

A number of general preprocessing steps are followed when dealing with text from the field. Firstly, we remove all non-necessary HTML tags from WOM documents. Secondly, as a word may have different meanings when used in different parts of speech (POS), we use the Brill tagger [7] in a natural language toolkit (NLTK) to choose a suitable part of speech tag for that word. Thirdly, we lemmatize words to their base forms according to WordNet, as morphs usually present the same meaning. Fourthly, a stop list with 596 stop words is used for word cleansing. The preprocessing of sentences using POS and stop words is commonly used in the field. For example, Chaturvedi et al. [13] used POS and stop words in their work on multilingual subjectivity detection. As the purpose of this paper is to form domain-oriented sentiment lexicons, only words found in SentiWordNet are retained, and those completely neutral words whose sentiment values in both positive and negative orientations are zero are omitted.