Expert Systems With Applications 44 (2016) 138–146

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Emotion space model for classifying opinions in stock message board

Banghui Luo[a], Jianping Zeng[b,*], Jiangjiao Duan[c]

[a] School of Computer Science, Fudan University, Shanghai 200433, China
[b] Engineering Research Center of Cyber Security Auditing and Monitoring, Ministry of Education, Shanghai 200433, China
[c] Business School, University of Shanghai for Science and Technology, Shanghai 200093, China

ARTICLE INFO

Keywords:
Opinion classification
Emotion
Stock message

ABSTRACT

The online stock message is known to have impacts on the trend of the stock market. Understanding investor opinions in stock message boards is important, and the automatic classification of the investors' opinions is one of the key methods for the issue. Traditional opinion classification methods mainly use terms and their frequency, part of speech, rule of opinions and sentiment shifters. But semantic information is ignored in term selection, and it is also hard to find the complete rules. In this paper, based on the classification of human emotions proposed by Ekman, we extend the traditional positive–negative analysis to the six important emotion states to build an extremely low dimensional emotion space model (ESM). It enables the prediction of investors' emotions in public. Specifically, we use lexical semantic extension and correlation analysis methods to extend the scale of emotion words, which can capture more words with strong emotions for ad hoc domain, like network emotion symbols. We apply our ESM on messages of a famous stock message board TheLion. We also compare our model with traditional methods information gain and mutual information. The results show that ESM is not parameter sensitive. Besides, ESM is efficient for modeling sentiment classifying and can achieve higher classification accuracy than traditional ones.

1. Introduction

Time series analysis methods are often used to forecast the change of stock price. These methods just utilize the statistical characteristics in stock price and then several models, such as hidden Markov model are applied (Al, Alam, & Rahman, 2014; Chen & Weiyin, 2015), ignoring external factors that influence stock price. Hence, the prediction accuracy is limited. Therefore, finding new factor for stock price prediction is in urgent. Recently, with the popularity of online forums, and social media, many people tend to express their attitudes on the stock price movements on the Internet (Duan, Wei, & Jianping, 2009; Kima & Kimb, 2014). Message on stock board represents successful attempts to manipulate stock prices and there is a strong association between Internet message board activity and abnormal stock returns and trading volume (Duan & Zeng, 2013; Kima & Kimb, 2014; Tumarkin & Whitelaw Robert, 2001). Opinion analysis provides new factors for predicting stock price. It is helpful improving the prediction accuracy (Duan et al., 2009; Si et al., 2013), and allowing regulators to understand the investors' sentiment in stock market (Sabherwal, Sarkar, & Zhang, 2011). Since humans are not able to process and interpret the large amounts of available data source, automated solutions are required, and the analysis of investors' opinion has attracted great attention.

Technically, the task of mining investors' opinions from stock message boards boils down to sentiment analysis of comment data—identifying and extracting different opinions from comments from investors. Although much work has been done recently on text mining (Achim, Altuntas, Häusser, & Kessler, 2011; Du & Rada, 2014; Ghiassi, Skinner, & Zimbra, 2013; Mukherjee & Bhattacharyya, 2012), most existing work aims at extracting and analyzing in a high dimensional space, and the parameters in these model are probably very sensitive (Ghiassi et al., 2013; Moraes, Francisco, & Gavião, 2013). Moreover, we do not know what the extract meaning the nonobjective space represents and the running time grows fast with higher dimension (Chen & Lazer, 2013).

In this paper, we propose a novel approach to convert the text into a low dimension emotion space ESM. Obviously, investors' emotion changes with their opinion on a stock. Investors' opinion can extract from their emotion in return. People are more likely to express a variety of emotions. As a consequence, we can extract investors' opinion through emotions. According to the research of notable psychologist Ekman Paul, people's emotions are classified into six basic categories, i.e., anger, disgust, fear, happiness, sadness, and surprise. Other emotions can be a combination of the six basic emotions. Certain emotions appeared to be universally recognized and independent from

[*] Corresponding author. Tel.: +86 13564317273.
E-mail addresses: bluo13@fudan.edn.cn (B. Luo), zjp@fudan.edu.cn (J. Zeng), jjduan@usst.edu.cn (J. Duan).

http://dx.doi.org/10.1016/j.eswa.2015.08.023
0957-4174/© 2015 Elsevier Ltd. All rights reserved.

cultural environment (Ekman & Friesen, 1971). Hence, it serves as a solid theory for us to construct ESM. Six important emotion states are used to enable the prediction of mood in general public (Chen & Lazer, 2013; Zhao, Dong, & Wu, 2012). The result of experiments on a number of tweets posted on Twitter shows it is believed that the six general state of emotion can be predicted with statistical significance.

The basic idea of the approach is to annotate a small size of words, which have definite and clear emotional tendency with emotion labels. Specifically, in extending emotion words based on the labeled ones, we use two approaches to weigh words by emotion tags. Then, message documents are mapped into the emotional space by a computation of the total weight of each emotion category in the message. Finally, classifying methods can be utilized to get the opinion label for unknown messages.

The main contributions of the paper are as follows. First, we propose a novel and extremely low dimensional emotion space model ESM, which is based on Ekman psychology theory for opinion classification. Second, we design two approaches to extend the number of emotion words, and vectoring the messages into emotion space with effective weight computation. Finally, experiments are done on the famous TheLion stock forum. The effectiveness confirms the feasibility of our approaches.

The proposed approach is quite general and has many potential applications. The mining results are quite useful for summarizing search results, monitoring public opinions, predicting user behaviors, and making business decisions. Our method requires no prior knowledge about a domain, and can extract general sentiment models applicable to any ad hoc queries. Although we only tested the ESM on stock message board, it is applicable to any text data with mixed domains and sentiments, such as customer reviews and films criticism.

The rest of this paper is organized as follows. Related work on opinion classification is described in the next section. Then the proposed methods including problem formulation, building emotion space model, extending labeled emotion words, are described in the third section. In the fourth section, we conduct the experiment and analyze the results. Finally, the conclusion is drawn.

2. Related work

In this section, we describe related research field of vector space model (VSM) and sentiment feature selection for sentiment classifying. We also discuss some problems in sentiment classifying domain.

2.1. Vector space model

Several kinds of methods can be employed to perform opinion classification. Opinion classification is traditionally treated as text classification. Hence, VSM is usually selected as the representation of text. Each dimensional word can be selected by some feature selection methods. (Deng, Kunhu, & Hongliang, 2014) conducts a study about several term weighting methods. These statistical functions include document frequency (DF), information gain (IG), mutual information (MI), odd ratio (OR). Support vector machines methods applied for testing different domains of data sets and using several weighting schemes for opinion classifying have been summarized (Moraes et al., 2013; O'Keefe & Koprinska, 2009; Saleh, Martín-Valdivia, Montejo-Ráez, and Ureña-López, 2011). They systematically evaluate a range of feature selectors and feature weights with both Naïve Bayes and Support Vector Machine classifiers. It is found that standard machine learning techniques definitively outperform human-produced baselines (Pang, Lee, & Vaithyanathan, 2002). Peñalver-Martinez et al., (2014) use ontologies at the feature selection stage to improve feature-based opinion mining. They take advantage of new Semantic Web-guided solutions to enhance the results obtained with traditional sentiment analysis processes, improving the vector space analysis methods.

One of the most commonly used methods is Information Gain (Ahmed, Hsinchun, Chen, & Salem, 2008; O'Keefe & Koprinska, 2009; Tan & Zhang, 2008). In general terms, the expected information gain is the change in information entropy from a prior state to a state that takes some information. The mutual information is equal to the total entropy for an attribute if for each of the attribute values a unique classification can be made for the result attribute.

Chi-square test is also an effective way for feature selection (O'Keefe & Koprinska, 2009; Tan & Zhang, 2008). Chi-squared test, also referred to as χ^2 test, is any statistical hypothesis test in which the sampling distribution of the test statistic is a chi-squared distribution when the null hypothesis is true. Also considered a chi-squared test is a test in which this is asymptotically true, meaning that the sampling distribution can be made to approximate a chi-squared distribution as closely as desired by making the sample size large enough. The chi-square (I) test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories.

Mutual information is a measure of the variables' mutual dependence (O'Keefe & Koprinska, 2009). It is commonly used, too. Intuitively, mutual information measures the information that X and Y share: it measures how much knowing one of these variables reduces uncertainty about the other. For example, if X and Y are independent, then knowing X does not give any information about Y and vice versa, so their mutual information is zero.

In an empirical study of sentiment categorization, four feature selection methods, information gain, mutual information, CHI and document frequency are adopted, in their work (Deng et al., 2014; Gong, Jianping, & Shiyong, 2011; Tan & Zhang, 2008). The experimental results indicate that information gain performs the best for selecting the sentiment terms. A subsumption hierarchy to formally define different types of lexical features and their relationship to one another, both in terms of representational coverage and performance is proposed (Riloff, Patwardhan, & Wiebe, 2006). They show that the reduce feature set can improve the performance on three opinion classification tasks, especially when combined with traditional feature selections approaches. Otherwise, Fisher's discriminant ratio is applied to select features for subjectivity text sentiment classification (Wang, Deyu, Xiaolei, Yingjie, & Hongxia, 2011). It combines different feature selection methods with two kinds of candidate feature sets and works well. But such approaches strongly associated with word frequency exaggerate the role of the low-frequency words, a word is in a class of each article document appears only once; but is more important than the 99% in this article the document appeared 10 words, actually behind the words is more representative, but only because it appears the number of documents less than at the front of the word "1", the feature selection can screen out the back of the word while retaining the former.

However, VSM only take the differentiation ability into consideration. The scale of VSM is dependent on the training dataset. As a result, the classification performance might degrade if the percentage of unknown words in test documents is high. Classification models, such as Bayes, SVM, KNN, etc., can be utilized to VSM-based opinion labeling (Das & Chen, 2007; Moraes et al., 2013). Five classification algorithms are integrated into classifying messages on Yahoo stock message boards into one of the three types, i.e., bullish, bearish and neutral (Das & Chen, 2007).

2.2. Sentiment feature selection

Another commonly used approach to opinion classification concentrates on sentiment feature and ignores those words without sentiment expression. Sentiment feature selection is one of the critical issues.

In Das and Chen (2007), the nature of a word, i.e., noun, adjective, etc., are selected since the kind of words usually associate human