Journal of Informetrics 5 (2011) 313–322

Contents lists available at ScienceDirect

Journal of Informetrics

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

A neural network based approach for sentiment classification in the blogosphere

Long-Sheng Chen [a,*], Cheng-Hsiang Liu [b], Hui-Ju Chiu [a]

[a] Department of Information Management, Chaoyang University of Technology, 168 Jifong E. Rd., Wufong Township, Taichung County 41349, Taiwan, ROC
[b] Department of Industrial Management, National Pingtung University of Science and Technology, 1 Shuefu Road, Neipu, Pingtung, Taiwan, ROC

ARTICLE INFO

Article history:
Received 16 July 2009
Received in revised form 6 January 2011
Accepted 6 January 2011

Keywords:
Semantic orientation
Neural networks
Sentiment classification
Blogs

ABSTRACT

Recognizing emotion is extremely important for a text-based communication tool such as a blog. On commercial blogs, the evaluation comments by bloggers of a product can spread at an explosive rate in cyberspace, and negative comments could be very harmful to an enterprise. Lately, researchers have been paying much attention to sentiment classification. The goal is to efficiently identify the emotions of their customers to allow companies to respond in the appropriate manner to what customers have to say. Semantic orientation indexes and machine learning methods are usually employed to achieve this goal. Semantic orientation indexes do not have good performance, but they return results quickly. Machine learning techniques provide better classification accuracy, but require a lot of training time. In order to combine the advantages of these two methods, this study proposed a neural-network based approach. It uses semantic orientation indexes as inputs for the neural networks to determine the sentiments of the bloggers quickly and effectively. Several actual blogs are used to evaluate the effectiveness of our approach. The experimental results indicate that the proposed approach outperforms traditional approaches including other neural networks and several semantic orientation indexes.

1. Introduction

Blogs are one of the fastest growing sections of the emerging communication mechanisms (Cohen & Krishnamurthy, 2006; Lambiotte, Ausloos, & Thelwall, 2007; Singh, Veron-Jackson, & Cullinane, 2008; Tang, Tan, & Cheng, 2009). Bloggers record the daily events in their lives and express their opinions, feelings, and emotions in an on-line journal, or blog (Nardi, Schiano, Gumbrecht, & Swartz, 2004). One of the features of most blogs is the fact that readers can comment on-line on whatever the blogger wrote in his/her journal. This ability has facilitated interaction between bloggers and those that read their blog (Chau & Xu, 2007). In addition, many of these blogs contain reviews on many products, issues, etc. (Martin, 2005; Murphy, 2006; Tang et al., 2009). Some of these reviews or evaluations found in a blogger's log may be negative, and sometimes they spread like wildfire through cyberspace. Negative comments about a product can be harmful to an enterprise. It is therefore important to effectively recognize the sentiment of bloggers, especially for those enterprises that use blogs as a marketing channel (Singh et al., 2008).

Lately, researchers have been paying much attention to sentiment classification and analysis (Subasic & Huettner, 2001; Wu, Chuang, & Lin, 2006) which identifies the emotions of their customers to allow companies to respond in the appro-

* Corresponding author. Tel.: +886 4 2332 3000x7752; fax: +886 4 2330 4902.
E-mail addresses: lschen@cyut.edu.tw, lschen0227@gmail.com (L.-S. Chen).

1751-1577/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.joi.2011.01.003

priate manner to what customers have to say. Nowadays, sentiment analysis became an important subfield of the field of information management (Tang et al., 2009) and can provide commercial bloggers with a tool to estimate the extent of product acceptance and to determine strategies that might improve a product's quality (Prabowo & Thelwall, 2009). Another advantage of recognizing the emotional state of bloggers is that it enables the companies to adjust their type of response to bloggers with negative comments (Lee, Narayanan, & Pieraccini, 2002). Another researcher indicated that the ability to identify the emotional state can reduce translation data size thereby increasing the fluency of net conferencing (Boucouvalas, 2002). Sentiment classification has also been widely applied in areas such as product comparisons, opinion summarizing, and reason mining (Tang et al., 2009; Yanaru, 1995).

The most direct method for determining a blogger's emotions is to use emotional keywords from the blogger's blog (Zhuang, Jing, & Zhu, 2006). Although a blog often contains multimedia data, simple text remains its main communication tool (Wu et al., 2006). In the literature, two popular approaches, machine learning (ML) methods and information retrieval (IR) (Hearst, 1992) techniques (semantic orientation index, SO index) (Tang et al., 2009), are employed to address the issue (Chaovalit & Zhou, 2005). In ML, several approaches have been developed. For examples, Abbasi, Chen, Thoms, and Fu (2008) proposed the support vector regression correlation ensemble (SVRCE) approach to analyze emotional states. Pang, Lee, and Vaithyanathan (2002) investigated several supervised ML methods to semantically classify movie reviews. Turney (2002) employs a specific unsupervised learning method for the semantic orientation of a review classification. Dave, Lawrence, and Pennock (2003) developed a method for automatically classifying positive and negative reviews and experimented with several methods related to the selection of features and scoring. In the work of Chaovalit and Zhou (2005), they used ML methods and the SO index to classify the comments of movie reviewers. Prabowo and Thelwall (2009) used multiple classifiers in a hybrid manner which may be more effective than any one of the classifiers separately. However, they did not consider the issue of training time required. These experimental results indicate that ML techniques have a better performance than the IR methods, but they require more time to be trained.

In IR, association, Pointwise Mutual Information (PMI), and Latent Semantic Analysis (LSA) have been employed to measure the similarity between words to classify sentiments. Several works reported that IR techniques are still good tools for sentiment classification. For examples, Turney and Littman (2003) employed association to determine the semantic orientation. Devillers, Luniel, and Vasilescu (2003) measured the emotional state by computing the conditional probability of keywords and related emotional states. Tao and Tan (2004) evaluated emotional conditions by utilizing emotive function words instead of emotive keywords. Hu and Liu (2004) used the adjective synonym sets and antonym sets in WordNet to judge the semantic orientation of adjectives. The classification accuracy of IR techniques is not very high (Pang et al., 2002); however, it allows us to quickly obtain the results of sentiment classification. In addition to the IR methods and ML techniques, there are various other ways to classify sentiments. Subasic and Huettner (2001) use fuzzy logic to manually construct a lexicon, based on which fuzzy technique was applied to which fuzzy set to analyze the affect of a document. Natural language processing (NLP) techniques are also used in this area (Tang et al., 2009).

The above methods either require a certain amount of manual construction or they must rely on externally structured information sources (Wang, Lu, & Zhang, 2005). Although ML techniques have better classification abilities, they require additional learning time. Besides, the information of classes which has to be determined by domain experts must be provided before training (Pang et al., 2002; Su, Chen, & Chiang, 2006). Therefore, so as to avoid the problems of both these methods but keep their strengths, this study proposed a neural network (NN) based approach which combines the advantages of the ML techniques and the IR techniques. In our proposed method, the back-propagation neural network (BPN) (Rumelhart & McClelland, 1986) has been selected as the basic learner based on its strength of fault tolerance. Our method uses the results of the SO indexes as the inputs for the BPN. Several cases collected from real world blogs or databases are provided to demonstrate the effectiveness of our method. The experimental results indicate that our method can efficiently increase the performance of sentiment classification and save a substantial amount of training time compared with traditional IR and ML techniques, respectively.

2. Neural networks and semantic orientation indexes

The proposed approach uses 4 different types of SO indexes as the input neurons. A brief introduction regarding BPN and the 4 different types of SO indexes is provided in the following subsections.

2.1. Back-propagation neural networks

Neural networks offer advantages such as adaptive learning, parallelism, fault tolerance, and generalization. In general, neural nets can be classified into two categories, feed-forward and feedback networks. In this study, the feed-forward network was employed because of its superior classification ability.

Among the feed-forward networks, BPN is the best known networks and it remains one of the most useful ones. This iterative gradient algorithm is designed to minimize the mean square error between the actual output of a multilayer feed-forward perceptron and the desired output. Based on the rule of thumb and reports of available published papers (Chen, Su, & Chen, 2009; Su, Hsu, & Tsai, 2002; Su, Yang, & Ke, 2002), the number of hidden layers could be one or two. The back-propagation algorithm includes a forward pass and a backward pass. The purpose of the forward pass is to obtain the activation value, and the backward pass is to adjust the weights and biases based on the difference between the desired