Chapter 42
Integrating Feature and Instance Selection Techniques in Opinion Mining

Zi-Hung You
Department of Nephrology, Chiayi Branch, Taichung Veterans General Hospital, Chiayi, Taiwan

Ya-Han Hu
https://orcid.org/0000-0002-3285-2983
Department of Information Management, National Central University, Taoyuan, Taiwan & Center for Innovative Research on Aging Society (CIRAS), Chiayi, National Chung Cheng University, Taiwan & MOST AI Biomedical Research Center at National Cheng Kung University, Tainan, Taiwan

Chih-Fong Tsai
Department of Information Management, National Central University, Taiwan

Yen-Ming Kuo
Department of Information Management, National Chung Cheng University, Chiayi, Taiwan

ABSTRACT

Opinion mining focuses on extracting polarity information from texts. For textual term representation, different feature selection methods, e.g. term frequency (TF) or term frequency–inverse document frequency (TF–IDF), can yield diverse numbers of text features. In text classification, however, a selected training set may contain noisy documents (or outliers), which can degrade the classification performance. To solve this problem, instance selection can be adopted to filter out unrepresentative training documents. Therefore, this article investigates the opinion mining performance associated with feature and instance selection steps simultaneously. Two combination processes based on performing feature selection and instance selection in different orders, were compared. Specifically, two feature selection methods, namely TF and TF–IDF, and two instance selection methods, namely DROP3 and IB3, were employed for comparison. The experimental results by using three Twitter datasets to develop sentiment classifiers showed that TF–IDF followed by DROP3 performs the best.

DOI: 10.4018/978-1-6684-6303-1.ch042

1. INTRODUCTION

The prevalence of computing and mobile technologies and the proliferation of social media such as Twitter, Facebook, and Google+ have extended the means of human communication (Ellison, 2007; Xiaomei et al., 2018). According to Thelwall et al. (2011), Twitter has more than five hundred million registered users, and more than 80% of them record their daily experiences and events on the Internet. Consequently, certain businesses have paid considerable attention to customer satisfaction with their products and services. In particular, opinions or comments that are based on text messages can be further analyzed.

Opinion mining or sentiment analysis poses a critical research problem (Li and Wu, 2010; Bravo-Marquez et al., 2013; Hu et al., 2017; Kaue & Moreira, 2016; Khan et al., 2014). It is usually based on text mining techniques including natural language processing (NLP), text analysis, and computational linguistics for identifying and extracting meaningful information from online opinions and reviews, news, or other sentences to predict the emotional state of users (Chatterjee et al., 2018; Deshmukh & Tripathy, 2018; Hu et al., 2018; Lee et al., 2018; Nguyen & Nguyen, 2018; Piryani et al., 2017; Tian et al., 2014). In general, useful keywords or terms related to emotions are extracted on the basis of NLP algorithms (Manning and Schutze, 1999) and publicly available resources such as SentiWordNet[1]; conversely, constructed sentiment classifiers can be used to classify the valence of certain selected opinions.

Opinion mining or sentiment analysis is similar to text classification or categorization. However, effective opinion mining poses two challenges. The first one is that the restrictions of writing an opinion differ substantially between different social network websites, which renders extracting the content of an opinion difficult. Consider, for example, Twitter; a single message is restricted to 140 characters. Under this constraint, fully describing a user's opinion can be extremely difficult.

Researchers in related works have attempted to extract representative keywords to describe sentimental statuses (Bravo-Marquez et al., 2013; Hu et al., 2017; Khan et al., 2014; Li & Wu, 2010; Nguyen & Nguyen, 2018; Tian et al., 2014). This feature extraction step can produce representative features for each opinion and thus enable constructing a classifier. However, the number of features extracted to describe an opinion (or text message) is usually very high, resulting in very high-dimensional feature vectors. Certain features in the high-dimensional feature vectors are not beneficial for the classification task, and such features can be considered noisy information. Therefore, performing feature selection to filter unrepresentative features has been widely studied in research on text classification (Aghdam et al., 2009; Lee & Lee, 2006; Rehman et al., 2017; Zheng et al., 2004). Recently, this topic has been considered in studies on opinion mining (Abbasi et al., 2008; Agarwal & Mittal, 2013; Nicholls & Song, 2010), but it has not been fully explored.

The second challenge is that when the volume of text documents is certainly large, certain documents can generally be considered outliers, which may degrade the final classification performance. Performing instance selection for filtering unrepresentative data has been shown to be effective in enhancing the performance of a classifier relative to a classifier without instance selection (Li & Jacob, 2008; Wilson & Martinez, 2000). Recent related studies have considered instance selection in text classification (Tsai & Chang, 2013; Tsai et al., 2014; Vinodhini & Chandrasekaran, 2017). However, few studies have focused on the effect of performing instance selection on opinion mining.

The research objectives of this study were (1) to provide a comprehensive evaluation of performing both feature and instance selections in opinion mining and (2) to develop a novel opinion mining process by integrating feature and instance selection techniques.