Information Sciences 181 (2011) 1138–1152

Contents lists available at ScienceDirect

Information Sciences

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

Ensemble of feature sets and classification algorithms for sentiment classification

Rui Xia [a,*], Chengqing Zong [a], Shoushan Li [b]

[a] National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China
[b] Department of Computer Science and Technology, Soochow University, Suzhou 215006, China

ARTICLE INFO

Article history:
Received 28 December 2009
Received in revised form 9 October 2010
Accepted 18 November 2010

Keywords:
Sentiment classification
Text classification
Ensemble learning
Classifier combination
Comparative study

ABSTRACT

In this paper, we make a comparative study of the effectiveness of ensemble technique for sentiment classification. The ensemble framework is applied to sentiment classification tasks, with the aim of efficiently integrating different feature sets and classification algorithms to synthesize a more accurate classification procedure. First, two types of feature sets are designed for sentiment classification, namely the part-of-speech based feature sets and the word-relation based feature sets. Second, three well-known text classification algorithms, namely naïve Bayes, maximum entropy and support vector machines, are employed as base-classifiers for each of the feature sets. Third, three types of ensemble methods, namely the fixed combination, weighted combination and meta-classifier combination, are evaluated for three ensemble strategies. A wide range of comparative experiments are conducted on five widely-used datasets in sentiment classification. Finally, some in-depth discussion is presented and conclusions are drawn about the effectiveness of ensemble technique for sentiment classification.

1. Introduction

Text classification has been one of the key tools to automatically handle and organize text information for decades. In recent years, with more and more subjective information appearing on the internet, sentiment classification [21,29], as a special case of text classification for subjective texts, is becoming a hotspot in many research fields, including natural language processing (NLP), data mining (DM) and information retrieval (IR).

The dominant techniques in sentiment classification generally follow traditional topical text classification approaches, where a document is regarded as a bag of words (BOW), mapped into a feature vector, and then classified by machine learning techniques such as naïve Bayes (NB) [19], maximum entropy (ME) [27], or support vector machines (SVM) [14]. The effectiveness of machine learning techniques when applied to sentiment classification tasks is evaluated in the pioneering research by Pang et al. [30]. The experimental results on the movie-review dataset produced via NB, ME, and SVM are substantially better than those results obtained through human generated baselines. But their performance is not as remarkable as when they are used in topical text classification. The main reason may be that traditional BOW does not capture word order information, syntactic structures and semantic relationships between words, which are essential attributes for sentiment analysis. Therefore, various kinds of feature sets, such as part-of-speech (POS) based features [12], higher-order *n*-grams [5,15,30], word pairs and dependency relations [5,10,15,36], have been exploited to improve sentiment classification performance.

Previous work, however, mostly focuses on joint features while ignoring an efficient integration of different types of features to enhance the sentiment classification performance. On one hand, among different classification algorithms, which

* Corresponding author.
E-mail addresses: rxia@nlpr.ia.ac.cn (R. Xia), cqzong@nlpr.ia.ac.cn (C. Zong), shoushan.li@gmail.com (S. Li).

0020-0255/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2010.11.023

one performs consistently better than the others remains a matter of some debate. On the other hand, different types of features have distinct distributions, and therefore would probably vary in performance between different machine learning algorithms. For example, it is reported in [30] that on the movie dataset, SVM performs the best, ME maintains an average, and NB tends to do the worst on unigram features; while the outcome is the reversed for bigrams. This is possibly due to the relevance between bigrams being lower than between unigrams. Moreover, the performance of classification algorithms is also domain-dependent. For instance, subsequent literature [4] shows that, using the same unigram features, NB performs better than SVM on datasets other than movie reviews.

We therefore intuitively seek to integrate different types of features and classification algorithms in an efficient way in order to overcome their individual drawbacks and benefit from each other's merits, and finally enhance the sentiment classification performance.

The ensemble technique, which combines the outputs of several base classification models to form an integrated output, has become an effective classification method for many domains [13,17]. In topical text classification, several researchers have achieved improvements in classification accuracy via the ensemble technique. In the early work [18], a combination of different classification algorithms (*k*-NN, Relevance feedback and Bayesian classifier) produces better results than any single type of classifier. Literature [6] makes a comparison of several ensemble methods for text categorization, which investigates six homogeneous ensemble methods (*k*-fold partitioning, bagging, boost, biased *k*-partitioning, biased *k*-fold partition, and biased clustering). In the field of sentiment classification, however, related works are very rare and no extensive evaluation has been carried out. Literature [38] proposes four ensemble algorithms (bagging, boosting, random subspace, and bagging random subspaces) using SVM as the base classifier and reports that ensemble of random subspaces can increase classification accuracy and the bagging subspaces model has the highest accuracies. In [20], different classifiers are generated through training with different sets of features, then component classifiers are selected and combined using several fixed combination rules. Experimental results show that all of the combination approaches can outperform individual classifiers and the sum rule achieves the best performance.

In this paper, we aim to make an intensive study of the effectiveness of ensemble techniques for sentiment classification tasks. Rather than an ensemble of different data re-sampling methods (e.g. bagging and boosting), we focus on ensemble of feature sets and classification algorithms. We design two schemes of feature sets that are particular to sentiment analysis: one is part-of-speech (POS) based and the other is word-relation (WR) based. For each scheme, we utilize NB, ME, and SVM as the base-classifiers to predict classification scores. In the ensemble stage, we apply three types of ensemble method (fixed combination, weighted combination, and meta-classifier combination) with three ensemble strategies (ensemble of feature sets, ensemble of classification algorithms, and ensemble of both feature sets and classification algorithms). A wide range of comparative experiments are conducted on five datasets widely used in sentiment classification. We seek answers based on empirical evidence to the following questions:

(1) What are the strengths and weaknesses of existing feature sets and classification algorithms when applied to the task of sentiment classification?
(2) Can the performance of a sentiment classification system benefit from the ensemble technique? To what extent can each of the three ensemble strategies improve the system performance?
(3) Among various combination methods, which one can be selected as the winner across all settings and datasets? Are there any guidelines to help choose the best from these methods?

The remainder of this paper is organized as follows. Sections 2 and 3 review traditional sentiment feature engineering and classification algorithms, respectively. In Section 4, we describe two schemes of feature sets and present the ensemble framework for sentiment classification. Experimental results are presented and analyzed in Section 5. In Section 6, we make in-depth discussion and answer the above three questions. Section 7 draws conclusions and outlines directions for future work.

2. Feature engineering

The text representation method dominating the literature is known as the BOW framework. In this framework, a document is considered as a bag of words and represented by a feature vector containing all the words appearing in the corpus. Although BOW is simple and quite efficient in text classification, a great deal of the information from the original document is discarded, word order is disrupted, and syntactic structures are broken. Therefore, sophisticated feature extraction methods with a deeper understanding of the documents are required for sentiment classification tasks. Instead of using a bag of words (unigrams), alternative ways to represent text, including POS based features, higher-order *n*-grams, and word dependency relations are presented in the literature.

2.1. Part-of-speech information

POS information is supposed to be a significant indicator of sentiment expression. The work on subjectivity detection [12] reveals a high correlation between the presence of adjectives and sentence subjectivity, yet this should not be taken to mean