Neurocomputing 167 (2015) 434–442

Contents lists available at ScienceDirect

Neurocomputing

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

New term weighting schemes with combination of multiple classifiers for sentiment analysis

Mohamed Abdel Fattah [a,b,*]

[a] Department of Computer Sciences, CCSE, Taibah University, Saudi Arabia
[b] Department of Electronics Technology, FIE, Helwan University, Cairo, Egypt

ARTICLE INFO

Article history:
Received 5 November 2014
Received in revised form
22 February 2015
Accepted 16 April 2015
Communicated by Y. Chang
Available online 29 April 2015

Keywords:
Sentiment classification
Openion mining
Term weighting schemes

ABSTRACT

The rapid growth of social media on the Web, such as forum discussions, reviews, blogs, micro-blogs, social networks and Twitter has created huge volume of opinionated data in digital forms. Therefore, last decade showed growth of sentiment analysis task to be one of the most active research areas in natural language processing. In this work, the problem of classifying documents based on overall sentiment is investigated. The main goal of this work is to present comprehensive investigation of different proposed new term weighting schemes for sentiment classification. The proposed new term weighting schemes exploit the class space density based on the class distribution in the whole documents set as well as in the class documents set. The proposed approaches provide positive discrimination on frequent and infrequent terms. We have compared our new term weighting schemes with traditional and state of art term weighting schemes. Some of our proposed terms weighting schemes outperform the traditional and state of art term weighting schemes results.

1. Introduction

Recent years have seen rapid growth in review sites and on-line discussion groups (e.g., the New York Times' Books web page) where a crucial characteristic of the posted articles is their opinion or overall sentiment towards the subject matter—for example, whether a product review is negative or positive [1–3]. Sentiment analysis is exploited to analyze people's opinions, evaluations, sentiments, attitudes, appraisals and emotions towards certain entities as services, products, individuals, organizations, events, topics and issues.

Sentiment classification is considered as a text classification task with two classes (positive and negative classes). Therefore, many existing supervised learning approaches can be exploited for this purpose. For instance, Naïve Bayes, Neural Networks, Hidden Markov Models, Gaussian Mixture Models and support vector Models. Many features and learning algorithms were exploited to train and test the supervised learning approaches such as: terms and their frequency, part-of-speech (POS), sentiment words and phrases, rules of opinions, sentiment shifters and syntactic dependency. Unsupervised Learning approaches can be exploited for sentiment classification task as well.

The dominant approaches in sentiment categorization generally follow traditional topical text categorization approaches [4], where a document is regarded as a bag of words (BOW), mapped into a feature vector, and then classified by machine learning techniques [5–11] such as naive Bayes (NB) [12], maximum entropy (ME) [13], or support vector machines (SVM) [9,14]. The effectiveness of machine learning techniques when applied to sentiment classification tasks is evaluated in the research by Pang et al. [15].

Yu and Hatzivassiloglou used a modified log-likelihood score value to determine the negative or positive orientation for each adverb, adjective, verb and noun [16]. Hu and Liu proposed a lexicon-based algorithm for sentiment orientation of a sentence [17]. The approach is based on a sentiment lexicon created using bootstrapping with given negative and positive sentiment word seeds and the antonyms and synonyms relations in WordNet. Kim and Hovy used a similar approach. However, they calculated the sentiment orientation by multiplying the scores of the sentiment words in the sentence [18]. Kim et al. exploited supervised learning methods. Some approaches used a domain specific lexicon with a shallow natural language processing approach to specify the sentiment orientation of a sentence [19–21]. Anthony and Gamon exploited a semi-supervised learning algorithm which is based on Expectation Maximization (EM) using naive Bayes as a

* Corresponding author at: Department of Electronics Technology, FIE, Helwan University, Cairo, Egypt.
E-mail address: mohafi2003@helwan.edu.eg

http://dx.doi.org/10.1016/j.neucom.2015.04.051
0925-2312/

classifier to learn from a small set of labeled sentences and a large set of unlabeled sentences [22]. Ryan et al. proposed a hierarchical sequence learning algorithm similar to conditional random fields (CRF) to learn sentiment [23,24]. Täckström and McDonald proposed an integrated supervised and a partially supervised model to perform multi-level sentiment classification [25]. Hassan, Qazvinian and Radev proposed a method to specify attitudes about participants in online discussions [26]. Davidov, Tsur and Rappoport proposed a supervised learning approach for sentiment classification of Twitter postings [27]. Zhang et al. proposed a sentiment elicitation approach that uses compositional semantic rule algorithm, bag-of-words with rule-based algorithm and numeric sentiment identification algorithm to train machine learning tool for classifying a tweet [28]. Liu et al. have investigated whether a flip polarity (switch negation) is a reasonable way to quantify negation or not [29]. Flip polarity seems to work well in certain cases and fails in others.

Vector Space Model (VSM) is exploited to represent documents for sentiment analysis task. The weight of each term in a document's vector is the key component of the VSM of document representation that measures the importance of the term in a document. In the indexing process, two features are of main concern: statistical term weighting where term weighting is based on discriminative supremacy of a term that appears in a document or a group of documents and semantic term weighting where term weighting is based on a term's meaning [30].

In information retrieval, complex term weighting approaches that are based on learning term weight by optimization are considered [31]. Although the complex term weighting approaches used in information retrieval task are proved to achieve good accuracy, the work of Paltoglou et al. showed that these approaches only provide information about the general distribution of terms without providing any evidence of class preference [32]. In Salton and Buckley's work, three factors have been considered for term weighting; term frequency, inverse document frequency and normalization [33]. Based on three factors, many approaches of term weighting have been presented for information retrieval. tfidf and normalized tfidf is considered as the best document weighting functions for information retrieval and text categorization tasks [34]. Although tfidf score gives positive discrimination to rare terms and is biased against frequent terms, it ignores the category information in text categorization task. Theeramnukkong and Lertnattee conducted different experiments based on various combinations of inter-class standard deviation, standard deviation with tfidf and class standard deviation [35]. Although the average results were bad compared with tfidf, one of their proposed approaches performed better than tfidf. In [36], idf has been replaced with information gain (X^2) statistic. However, it has not shown a consistent superiority over the standard tfidf. In [37], idf has been replaced with some feature parameter functions in text classification task. This work reported that tf.chi is better than tfidf when it is associated with SVM as a classifier.

A variety of feature selection techniques [46], such as information gain [38], chi-square test, and document frequency [39], have been used to reduce the dimension of the vectors. Soucy and Mineau introduced a new term weighting method (ConfWeight) based on statistical confidence intervals [40]. The experimental results of this work showed that ConfWeight outperformed tfidf when it was applied on three document sets. Lan et al. proposed a supervised term weighting approach, tf.rf, to improve the terms' discriminating power for text classification task [41]. The experimental results show that tf.rf gives better performance than some supervised term weighting approaches including tfidf. Luo et al. replaced the idf function with a semantic weight (sw) using WordNet [42]. The experiment's results showed that the proposed tf.sw scheme gives better results than tfidf scheme. Using semantic term weighting, it is possible to address a limited number of terms in a term index. However, it is difficult to provide the appropriate semantic knowledge of a term based on categories with large number of terms in the term index. Therefore, it is more convenient to exploit statistical term weighting approaches to compute the score of a term.

In this work, term frequency inverse document frequency, term frequency inverse class frequency, term weighting based on: mutual information, odds ratio, weighted log likelihood ratio and X^2 statistic are considered as base line approaches. In this paper, we propose new term weighting schemes for sentiment analysis. The proposed new term weighting schemes exploit the class space density based on the class distribution in the whole document set as well as the class documents set. These proposed approaches provide positive discrimination on frequent and infrequent terms. Support vector machine classifier (SVM), probabilistic neural network (PNN), Gaussian mixture model (GMM) have been exploited to investigate the traditional and proposed term weighting schemes effectiveness. Combination of the previously mentioned classifiers using simple voting and Borda count approaches have been exploited as well.

The paper is organized as follows: Section 2 illustrates the proposed term weighting schemes with combination of multiple classifiers, Section 3 describes the experimental results and Section 4 is the conclusions and future works.

2. The proposed term weighting schemes with combination of multiple classifiers

Before we present our proposed term weighting schemes, we investigate some traditional term weighting schemes to exploit them as base line approaches.

2.1. Traditional term weighting schemes

2.1.1. Term frequency inverse document frequency (tfidf)

Given a set of classes $C=\{c_p$ (positive class), c_n (negative class)$\}$ and a set of training documents $D=\{d_1, d_2,\ldots d_N\}$ where each training document d_i is assigned to one class (positive or negative), sentiment classification is a task to use this given information to find one suitable category for a new document. In a vector space model, a document is represented by a vector based on the weight of each term in the document.

Term frequency (tf) and inverse document frequency (idf) in the form of tfidf is applied in most research works to weight a term in a document. The term frequency in the given document is simply the number of times a given term appears in that document. This count is usually normalized to prevent a bias towards longer documents (which may have a higher term count regardless of the actual importance of that term in the document) to give a measure of the importance of the term t_i within the particular document d_j. Thus the term frequency is defined as follows:

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

where $n_{i,j}$ is the number of occurrences of the considered term in document d_j, and the denominator is the sum of number of occurrences of all terms in document d_j. The inverse document frequency is a measure of the general importance of the term (obtained by dividing the number of all documents by the number of documents containing the term, and then taking the logarithm of that quotient).

$$\mathrm{idf}_i = \log \frac{|D|}{|\{d : t_i \in d\}|} \tag{2}$$