Available online at www.sciencedirect.com

ScienceDirect

Karbala International Journal of Modern Science 1 (2015) 26–31
http://www.journals.elsevier.com/karbala-international-journal-of-modern-science/

Statistically weighted reviews to enhance sentiment classification

S. Prakash [a,*], T. Chakravarthy [a], E. Kaveri [b]

[a] *Dept of Computer Science, AVVM Sri Pushpam College, Poondi, Tamil Nadu, India*
[b] *Dept of Computer Science, Bharathidasan University Constituent College for Women, Orathanadu, Tamil Nadu, India*

Received 11 April 2015; revised 28 June 2015; accepted 9 July 2015
Available online 3 September 2015

Abstract

The exponential growth of Internet content, due to social networks, blogs and forums necessitate the research of processing the information in a meaningful way. The research area, Opinion mining is at the cross roads of computation linguistic, machine learning and data mining, which analyze the shared online reviews. Reviews may be about a product, service, events or even a person. Word weighting is a technique that provides weights to words in these reviews to enhance the performance of opinion mining. This study proposes a supervised word weighting method that combined, Word Weighting (WW) and Sentiment Weighting (SW). For WW and SW two function each applied based on word frequency. So totally four statistical functions are applied and checked on categorical labels. Support Vector Machine is used to classify the weighted reviews and it outperforms the existing weighting methods. Two different sizes of corpus are used for the verification.

© 2015 The Authors. Production and hosting by Elsevier B.V. on behalf of University of Kerbala. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Keywords: Sentiment analysis; Word weighting; Classification; Support vector machine; Accuracy

1. Introduction

The information shared in social network blogs and forums contain the healthy information about products, event service or popular persons. Opinion mining or sentiment analysis is the new area in which, these information are processed to aid decision for the customers and business people.

The researchers use methods which exist in text processing, machine learning and natural language processing [1–3]. Before classifying the reviews whether the user said about the product positive or negative, the text should be weighted, usually by binary weights [4] and also frequency based [5,6]. Some of the statistical methods are used as feature selection techniques to reduce the dimension [7] and weighting the feature [8]. The proposed method uses two variants of term frequency formula and two statistical methods for finding document frequency, in total there are four combination of weighting methods. Then to evaluate the effectiveness of these methods Support Vector Machine is used to check the weighting influence on classifier. The proposed methods provide best accuracy on bench mark data sets compared with basic tf.idf and BM25 weighting methods.

* Corresponding author.
E-mail addresses: prakashselvakumar@gmail.com (S. Prakash), tcvarthy@gmail.com (T. Chakravarthy), rpkaveri@gmail.com (E. Kaveri).

Peer review under responsibility of University of Kerbala.

http://dx.doi.org/10.1016/j.kijoms.2015.07.001
2405-609X/© 2015 The Authors. Production and hosting by Elsevier B.V. on behalf of University of Kerbala. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1.1. Existing methods

Word weighting includes the computation of how much information a word associated to a document is giving relevant to the classes. Though there is no mathematical proof to the tf.idf (term frequency and inverse document frequency), but intuitively many researchers have proved the process [9,10]. Another weighting method, a variant of tf.idf is BM25 used in various studies to provide better results than tf.idf [6,10–12]. Some studies apply weights by selection methods using statistical formulas such as chi square [13], gain ratio, information gain [8]. They had the best result by using CHI in the place of idf on Reuters-21578 dataset, classification by SVM [13]. One of the authors developed a new statistical confidence interval weighting technique which gives more accuracy than tf.idf [14]. To improve the words' discriminating power, tf.rf is used as weighting formula [15]. Class indexing based weighting method computes multiplication of tf.idf with its inverse class space density frequency [16]. Earlier, Pang pointed out binary weights for binary unigram document is the best baseline weighting [4]. Keeping this in mind, BM25, the variant of tf.idf is tried for text classification and proved its efficiency [17]. Both supervised and unsupervised methods are used to learn word features that get semantic and sentiment content, but their results show tf.idf as better method than the proposed one [18]. With this motivation, the proposed weighting techniques take the term frequency variants for word frequency calculation and statistical formulas to get the importance of word to the class.

1.2. Proposed work flow

The role of proposed weighting techniques and its significant role in classification is given in Fig. 1. Online reviews about movies are taken as corpus to prove the performance of the proposed methods. The corpus is preprocessed as given below.

a) Case Folding: Converting the upper case into lower case letters, which is called cleaning the reviews in the documents.
b) Tokenizing: splitting the sentences into separate words of each document.
c) Indexing: Document identification number is created.

The preprocessed documents are weighted by the proposed multiplicative combination of weighting methods such as two Word Weighting and two Sentiment Weighting methods. The corpus is weighted using four combination of methods separately and given to the Support Vector Machine classifier. SVM learns a model from the labeled data set (Positive/Negative) and classifies the test data set which is given by excluding the labels. The classifier is evaluated based on precision, recall and F1 measure. To verify the proposed weighting techniques, the results are compared against a popular existing weighting technique BM25. For this verification, existing studies which used the same Cornel Movie corpus and did the classification using BM25 are analyzed.

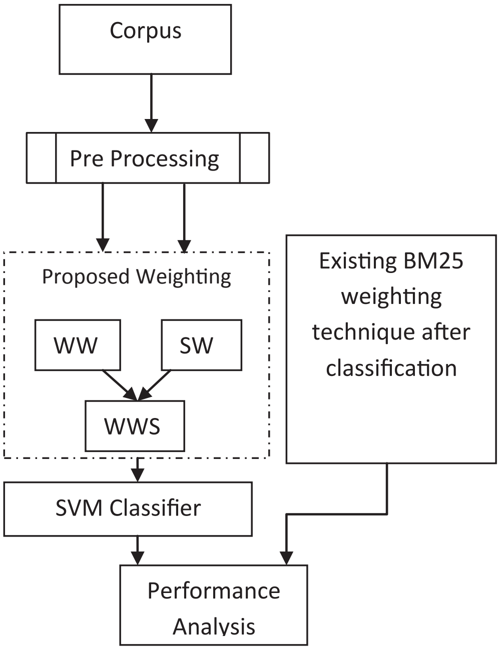

Fig. 1. Proposed work flow.

2. Statistical weighting scheme

Word Weighting to Sentiments (WWS) is computed by multiplying Word Weighting (WW) with Sentiment Weighting (SW). This study uses two variants of TF [19] as WW and two statistical formulas as SW.

2.1. Word weighting computation

Let the assumption of positive reviews be R^1 and set of negative reviews be R^2. Let $V = \{v_1, v_2, \ldots v_m\}$ is the unique word set of both review sets. Let document d_j contains word vector $d_j = (w_{1j}, w_{2j}, \ldots, w_{mj})$ and w_{ij} denotes weight of w_i in d_j. w_{ij} is computed as follows

Download English Version:

https://daneshyari.com/en/article/483927

Download Persian Version:

https://daneshyari.com/article/483927

Daneshyari.com