Chapter 3
Dimensionality Reduction Techniques for Text Mining

Neethu Akkarapatty
SCMS School of Engineering and Technology, India

Nisha S. Raj
SCMS School of Engineering and Technology, India

Anjaly Muralidharan
SCMS School of Engineering and Technology, India

Vinod P.
SCMS School of Engineering and Technology, India

ABSTRACT

Sentiment analysis is an emerging field, concerned with the analysis and understanding of human emotions from sentences. Sentiment analysis is the process used to determine the attitude/opinion/emotions expressed by a person about a specific topic based on natural language processing. Proliferation of social media such as blogs, Twitter, Facebook and Linkedin has fuelled interest in sentiment analysis. As the real time data is dynamic, the main focus of the chapter is to extract different categories of features and to analyze which category of attribute performs better. Moreover, classifying the document into positive and negative category with fewer misclassification rate is the primary investigation performed. The various approaches employed for feature selection involves TF-IDF, WET, Chi-Square and mRMR on benchmark dataset pertaining diverse domains.

INDRODUCTION

Mining is the process of extracting relevant information from large volume of data. The World Wide Web contains a huge volume of documents containing comments, feedback, critiques, reviews related to wide documents. Processing of natural language is a herculean task for humans to understand, analyze and to extract useful information from enormous amount of data. Thus the work helps to automatically determine the sentiment (positive or negative) of online texts is significant. Opinion mining or sentiment analysis (Liu, 2012), aim to extract the features upon which the reviewers express their opinions and help to determine whether the opinions are positive, negative or neutral.

DOI: 10.4018/978-1-5225-0489-4.ch003

In our day to day lives, analyzing the reviews/opinions has become an integral part for decision making. For example, if a person wishes to purchase a product online, he/she will refer to the prior reviews and comments posted by the experienced users in web. In order to enhance the product sales and to improve the customer's satisfaction, most of the on-line shopping sites provide facility for customers to write reviews/comments about the product they wish to purchase. But it seems to be a cumbersome task to read the entire reviews available in the web for purchasing a specific product. Hence, the user's interest is in determining if the reviews influences/ recommends in buying a product or not. If lot of reviews recommends buying the product, user will conclude to buy, otherwise not to buy (Feng, Zhang, & Deng, 2010).

Sentiment analysis has the wide spread applicative areas such as e-learning, automatic survey analysis, opinion extraction and recommender systems. In the past decade, opinion mining has been studied in fields like natural language processing, data mining, information retrieval, web mining etc.

SENTIMENT CLASSIFICATION

Sentiment classification is a part of opinion mining which refers to the task of extracting sentiment word from a given text and then classifying the content into positive or negative in its sentiment. Classification is usually performed at three levels namely:

1. Document.
2. Sentence.
3. Attribute level.

In the following section, each of them will be discussed in detail.

Document Level Sentiment Classification

Document level classification (Kumar, 2015; Bollegala, 2013; Singh, 2013; Mouthami, 2013; Wong, 2011) identifies the opinionated document (e.g product review) into classes such as positive, negative and neutral based on the overall sentiment expressed by the writer. The widely used dataset for document level sentiment classification is Cornell Movie review corpora (Mouthami, Devi & Bhaskaran, 2013) which were used in (Pang, 2004, 2002). Naïve Bayesian and Support Vector Machine are the supervised learning algorithms that are widely used to prepare model. The rating usually in the form of 1-5 stars is used by the reviewer for training as well as testing data. The features that are extracted can be any one/more combination of bag of words, adjectives from part of speech tagging, opinion words, phrases, negations, dependencies etc. Prior experiment results demonstrate that supervised learning is the most powerful method on preview of accuracy (Li & Liu, 2010). The unsupervised learning can also be performed by retrieving the opinion words inside a document. In order to find the semantics of the words that has been extracted, the point-wise mutual information can be used which in turn helps to improve the performance. The main challenge faced by the classification performed at the document level is that most of the sentences in a document seems to be irrelevant in expressing the opinion about an entity. As the comparative sentences appears in case of forums and blogs, customers compare one