CHAPTER 1

INTRODUCTION

1.1 SENTIMENT ANALYSIS

Sentiment detection is the process of finding the sentiment expressed in a review by using machine-learning techniques or natural language processing (NLP) techniques. This is also known as opinion mining (OM) and sentiment analysis. The objective is to classify a review document as either positive or negative based on the sentiment expressed in it. This type of classification is carried out at the document level, without discovering about what people liked or did not like.

Purchase decisions are greatly influenced by the opinion of the current users of the product. Nowadays the customers depend on opinion-rich information available online for any purchase decision. Since there is a huge demand for others opinions and sentiments, the information technologists have been researching on ways to uncover the sentiment expressed in the review documents. The companies are also keen on knowing whether their products/services are liked or not, in order to retain and increase their customer base. An automated sentiment analysis system detects not only the comments of individuals but also the overall opinions expressed.

1.2 SENTIMENT ANALYSIS METHODS

The research in sentiment analysis can be broadly classified into two categories viz. 1. Data mining approach 2. Natural language processing approach.

1.2.1 Data Mining Approach

Data mining approach accomplishes the job of classification by expressing the unstructured text documents as structured term-document matrix containing numerical scores. This approach borrows several techniques from computational linguistics and information retrieval to convert the unstructured data (text documents written in natural language) as structured data and after this the traditional data mining algorithms are applied for the purpose of classification.

Creation of the document-by-term frequency matrix is the first step in this approach. This is achieved by parsing each document into individual terms, or term/part-of-speech pairs. After this, the documents are represented as vectors of length equal to the number of terms in the entire collection. These vectors are sparse, containing mostly zeroes due to the fact that an individual document can contain only a small percentage of the terms in the entire collection.

As the matrix created during the first stage of this process is sparse, a dimensional reduction process is undertaken to represent each document on a reduced dimensional space containing the features that are significant for the purpose of classification. This is done by using the techniques like Singular Value Decomposition (SVD). During this step, the dimension of each document is reduced to just 50-100 features.

Finally, these reduced-dimensional vectors along with the sentiment label are supplied to the predictive model, which will learn from the patterns of the training data to create the predictive model. Sentiment expressed in document can be classified by using the predictive model created using the training set.

In spite of the advantages like discovering unimagined and complicated patterns, this approach requires large training data for its success. Also the vector representation is poor in distinguishing the meaning of the features, which have different meaning based on situations. For example, consider the following two phrases.

"Person for a great party".

"Great person for a party".

These two phrases convey two different meanings, but in a vector representation both the phrases would have identical representation, thus spoiling the classification accuracy.

1.2.2 Natural Language Processing Approach

Natural Language Processing (NLP) approach to sentiment analysis deals with automatic extraction of meaning/sentiment from the natural language text using POS tagging, developing a lexicon and pattern analysis. It is really a tough task to teach a machine to understand and think like a human. For example, consider the following review:

"….You can quibble about its clichés, predictability, and rare moments of overcooked sappiness, but none of that takes away from the entertainment value."

This review would certainly be interpreted as positive by humans, but it is very challenging to get the same output from a computer due to the presence of strong negative words.

The rule-based NLP methods use the syntactic patterns in the text along with certain entities to understand the meaning of the given text. A

combination of parts of speech, linguistic dictionaries, and noun phrases with a range of operators is used for the purpose of extracting the meaning. The commercially available software product like SAS contains the following operators:

1. Boolean operators – used to include or exclude various entities (e.g., AND, OR, NOT).
2. Frequency operators – used to count the number of occurrences of certain entities (e.g., MIN, MINOC, MAXOC)
3. Context operators – used to fix the context within which certain entities occur in the document.
4. Sequence operators – used to look for some specified sequence of entities.

1.3 MACHINE LEARNING ALGORITHMS

In the Machine Learning (ML) approach to sentiment analysis, the classifier automatically learns the properties of categories from the pre-classified training documents. ML based classification is called as supervised learning because this process is guided by the labeled training set. The points to be considered while using the ML classifiers are:

a) Decide on the categories that will be used to classify the instances.

b) Provide a training set for each of the categories.

c) Decide the number of features to represent the instances.

d) Decide the algorithm to be used for classification.

1.3.1 The General Structure of the Problem

According to Ronen Feldman and James Sanger (2009), the approximating function M: D x C⟶ {0, 1} is called a classifier and the task is to develop a classifier that delivers results as close as possible to the true category assignment function F: D x C⟶ {0, 1}, where D is the set of all documents and C is the predefined set of categories. F (d, c) assumes the value '1' if the document 'd' belongs to the category 'c' and 0 otherwise. Lewis (2000) and Sebastini (2002) discuss the general introduction about the text classification in detail.

1.3.2 Probabilistic Classifiers –The Naive Bayes Algorithm

Probabilistic classifiers use the Bayes' theorem to calculate the probability P (c|d), that a document belongs to a category c.

$$P(c|d) = P(d|c)\, P(c) / P(d) \qquad (1.1)$$

In order to determine the probability P (d|c), it is assumed that the coordinates representing the document as a feature vector are independent. These classifiers are called as Naïve Bayes (NB) classifiers. Some justification about the robustness of these classifiers can be found in Domingos and Pazzani (1997).

1.3.3 Bayesian Logistic Regression

Only in the recent past Bayesian Logistic Regression (BLR) has started gaining importance due to the high level of accuracy in text categorization. For the binary classification, the general form of the logistic regression model has the form

$$P(c|d) = \varphi(\beta.d) = \varphi\left(\sum_i \beta_i.d_i\right) \qquad (1.2)$$

Where c is the category membership and can take the value of "0" or "1", d = (d_1, d_2,….) is the document representation in the feature space, β = (β1, β2,….) is the model parameter vector and φ (x) is the logistic link function and is represented by,

$$\varphi(x) = \frac{\exp(x)}{1+\exp(x)} = \frac{1}{1+\exp(-x)} \tag{1.3}$$

Either a Gaussian or Laplace prior is used to describe the parameter vector β to assign a high probability to each β_i's being at or near zero.

1.3.4 Decision Tree Classifiers

A Decision Tree (DT) classifier resembles a tree in which the features are represented by nodes and the edges leaving a node are labeled by the feature weight and leaves represent the categories. The tree is constructed based on a recursive procedure. At each step a feature 'F' is picked and the training collection is divided into two groups, one containing 'F' and another not containing 'F'. This procedure is carried out until only documents of a single category remain. A leaf is generated at the end of this procedure. Information gain or entropy is used to choose a feature at each step. Generally the DT classifiers are used as baseline classifiers.

Apte et al (1994a, 1994b, 1994c), Li and Yamanishi (1999), Chen and Ho (2000) and Li and Yamanishi (2002) discuss the details on how to use the decision trees for text categorization.

1.3.5 Neural Networks

Neural network (NN) can be designed to carry out the task of opinion mining. The features of a document are the input nodes, the output nodes deliver the category. The dependence relations are taken care by the

link weights. Generally the NN are trained by back propagation, i.e. the training documents are fed into the input nodes and if a wrong classification occurs, the error is propagated back in the network to minimize the error by adjusting the link weights.

Perceptron is the simplest kind of a NN, which has only two layers viz. input layer and output layer. A multi layer perceptron contains one or more hidden layers between the input and output layers.

1.3.6 Regression Methods

A real-valued function can be approximated using the knowledge of its values on a set of points. This is the crux of regression techniques. The regression techniques can be used for Text Categorization (TC) and opinion mining problems if the assignment function is a member of a family of continuous real valued functions. Yang and Chute (1994) used the Linear Least-Square Fit (LLSF) regression technique for a TC problem. Zhang and Oles (2001), Zhang et al (2003), and Zhang and Yang (2003) discuss the text categorization using regression methods.

1.3.7 Support Vector Machines

The fastest and most widely used algorithm among all the ML algorithms is the support vector machine (SVM). A binary SVM is a hyperplane separating the feature space of positive instances from the feature space of negative instances. During the training phase, the hyperplane that can separate the positive feature space from the negative feature space with a maximal margin is chosen. The margin is distance of the nearest point from the positive and negative sets to the hyperplane. Support vectors, the subset of the training instances, determine the hyperplane for a SVM. SVM classifiers perform extremely well irrespective of the dimensionality of the feature space.

Joachims (1998), Joachims (1999), Joachims (2000), Joachims (2001), and Joachims (2002) discuss the details and approaches for using the SVM for text classification. Vapnik (1995) gives the complete theory of the statistical learning principles.

1.4 FUZZY LOGIC

Fuzzy logic based systems are handy in real life situations where the decision to be taken are based on multiple criteria with complex interlink among them. It is very true for a sentiment analysis process in which the system must be able to understand the sentiment expressed by a customer in a review based on the statements about various features of the product or service. For example, in a movie review, the reviewer may praise the director for the usage of technology and blame on the acting and story. Deciding on the overall sentiment as positive or negative depends on the opinion words or phrases used by the reviewer for each of the features. When the number of features is more, the complexity in the decision making gets added and hence the decision making becomes tough. In such situations, fuzzy logic can be effectively used.

A detailed analysis on the application of fuzzy logic for the sentiment analysis is given in chapter 5.

1.5 MAHALANOBIS DISTANCE BASED CLASSIFICATION

In 1930, P.C. Mahalanobis, founder of the Indian Statistical Institute, introduced a statistical measure called the Mahalanobis Distance (MD). MD is a superior statistical measure than the other statistical measures like Euclidean distance and Manhattan distance used for clustering and classification because it is based on the correlation among the various dimensions of the given problem (Genichi Taguchi and Rajesh Jugulum 2002).

Genichi Taguchi et al (2002) popularized the Mahalanobis-Taguchi System (MTS) and proposed a method for using the MD for the pattern recognition problems. They established that, for a pattern recognition problem, if a reference set can be created using the characteristic dimensions of the problem, then using the reference set, the test set can be classified either it belongs to the family of reference set or not by calculating the MD between the test set and the reference set. The reference set is called as Mahalanobis Space (MS).

For example, in a medical diagnosis system, the improvement after a medication can be checked by creating the MS using the data of a group of healthy people and then the MD between the MS and test data will depict any improvement has happened. It means, if the medication has resulted in a health improvement, then the MD between the patient's test data and MS will be less. If the medication has not improved the health, then the MD between the patient's test data and MS will be more.

1.5.1 Mahalanobis Distance

MATLAB, a software tool used for numerical computation and visualization, has built-in functions for calculating the MD. A detailed analysis on the application of MD for the sentiment analysis is explained in chapter 4.

1.6 ORGANIZATION OF THIS THESIS

Chapter 1 gives the general introduction to the sentiment analysis and the various approaches that are currently in use for the purpose of sentiment classification at document level.

Chapter 2 provides a detailed literature review on the field of sentiment analysis.

Chapter 3 provides the detailed procedure on representing the text documents in the form of a matrix using representative terms.

Chapter 4 provides the complete procedure on how MD can be used as a measure to detect sentiment expressed in document. The design and development of MD based classifier is explained in detail.

Chapter 5 provides details on how the concept of fuzzy logic can be used in sentiment analysis. Design and development of a fuzzy logic based classifier is explained in detail.

Chapter 6 provides details on the advantage of using fuzzy scores in the place of frequency of representative terms. Calculation of MD using fuzzy score of the representative terms is explained in detail.

Chapter 7 provides the complete picture of the hybrid classifier design.

Chapter 8 gives the comparative performance analysis of Mahalanobis distance based classifier (MDC), Mahalanobis-fuzzy system based classifier (MFSC) and fuzzy logic based classifier (FLC).

Chapter 9 gives the conclusion and the scope for future research work using the ideas proposed in this thesis.