Text Classification and Classifiers: A Comparative Study

[1]Payal R. Undhad, [2]Dharmesh J. Bhalodiya
[1]M.E. Student, [2]Assistant Professor
Dept. of Computer Engineering, BHGCET, Rajkot, India

***Abstract*-Text classification is used to organize documents in a predefined set of classes. It is very useful in Web content management, search engines; email filtering, etc. The expansion of information and power automatic classification of data and textual data gains increasingly and give high performance. In this paper some machine learning classifiers are described i.e. Naive Bayesian, KNN(K-nearest neighbor), SVM(Support Vector Machine), neural network. Which are classified the text data into pre define class. This paper surveys of text classification, process of text classification different term weighing methods and comparisons between different classification algorithms.**

Keywords: Text classification, KNN, Naïve bayes, Support Vector Machine, Decision Tree

1. INTRODUCTION

Text classification is a data mining technique used to predict categorical label. Aim of research on text classification is to improve the quality of text representation and develop high quality classifiers. Text classification process includes following steps i.e. collection of data documents, data preprocessing, Indexing, term weighing methods, classification algorithms and performance measure. Machine learning techniques have been actively explored for text classification. Machine learning algorithm for text classification are Naive bayes classifier [1], K-nearest neighbor classifiers [2], support vector machine [3] ,neural network. [4]

2. TEXT CLASSIFICATION PROCESS

Fig 1 represents the different stages of text classification which include collection of documents, preprocessing, indexing, feature selection, classification algorithm and performance measure. [4]

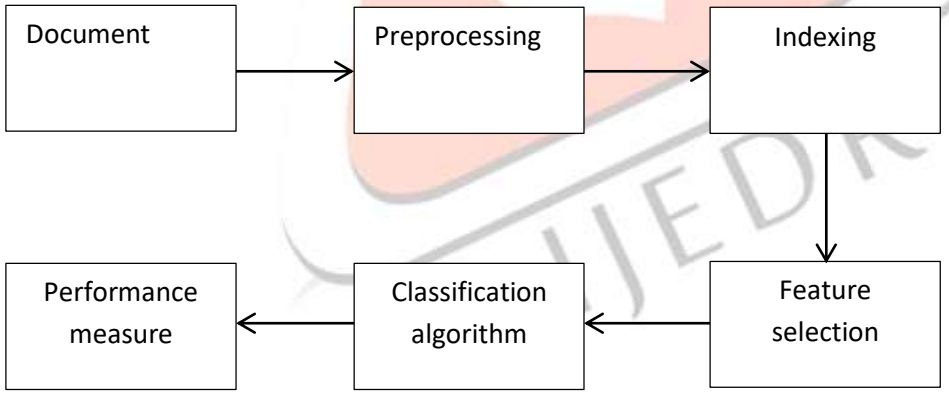

Fig 1. Stage of text classification [4]

2.1 Documents

Text classification process is start with the collection of the data from different type of format such as pdf, doc, html etc. large amount of data are available on the social media, and in various sources in different formats.

2.2 Preprocessing

Data mining is the process of extracting hidden or useful information from large dataset or document. Data of this document is often incomplete inconsistent and lacking in certain behavior and is likely to contain many errors. Data goes through a series of steps:

Removing stop words: Stop words are very common words that appear in every document they have little meaning, they serve only syntactic meaning but do not indicate subject matter it is well recognized among the conformation retrieval experts that a set of functional English words (eg. “the”, “a”, “and”, “that”, ”this”, ”is”, ”an”) is useless as indexing terms. These words have very low Discrimination value, since they occur in every English document [5]. Hence they do not help in distinguishing between documents about different topics. The process of removing the set of bearing functional words from the set of words produced by word extraction is known as stop words removal. In order to remove the stop words, first step is creating a list of stop words to be removed, which is also called as the stop word list. After this, second step is the set of words produced by word extraction is then scanned so that every word appearing in the stop list is removed.

Stemming: In stemming different forms of the same word are converted into a single word. For example, singular, plural, and different tenses are converted into a single word. Port stemmer algorithm is well-known algorithm for stemming e.g. connection to connect, computing to compute.

2.3 Indexing

The documents representation is one of the preprocessing technique which is used to reduce the complexity of the documents and make them easier to handle, the document have to be transformed from full text document to a document vector.

In current research mostly used Document representation is called vector space model(VSM) here documents are represented by vectors of words. Usually, one has a collection of documents which is represented by word by word document Matrix. VSM [6] document representation technique has its own disadvantages i.e. high dimensionality of the representation, loss of correlation with adjacent words and loss of semantic relationship that exist among the terms in a document to overcome these problems, term weighting methods are used to assign some predefine weights to the term.

2.4 Feature Selection

After completion of further steps the important step of text classification is feature selection [7] to construct vector space, which improves the accuracy of a text classifier. The main idea of Feature Selection is to select subset of features from the main documents. Feature selection is performed by keeping the words with highest score according to predetermined measure of the importance of the word. Because of for text classification a major problem is the high dimensionality of the feature space. Many feature evaluation metrics have been studied which are information gain (IG), term frequency, Chi-square, expected cross entropy, Odds Ratio, the weight of evidence of text, mutual information, Gini index. But Feature selection of association text mining is more efficient than IG and document frequency .Other various methods are presented like [58] sampling method which is randomly samples roughly features and then make matrix for classification. By considering problem of high dimensional problem [59] is presented new FS witch use the genetic algorithm (GA) optimization.

2.5 Classification

In current research automatic classification of documents into predefined categories has observed as an active attention, the documents can be classified in three ways, unsupervised, supervised and semi supervised methods. From last few years, the task of automatic text classification have been widely studied seems in this area, including the machine learning approaches such as Naïve bayes classifier, Decision Tree, K-nearest neighbor(KNN), Support Vector Machines(SVMs), Neural Networks. Some techniques are described in section 3 [8].

2.6 Performance Measure

Performance measure is Last stage of Text classification, in which the evaluations of text classifiers is typically conducted experimentally, rather than analytically. The experimental evaluation of classifiers, rather than concentrating on issues of Efficiency, usually tries to evaluate the effectiveness of a classifier, i.e. its capability of taking the right categorization decisions. An important issue of Text categorization is how to measures the performance of the classifiers. Many measures have been used, like Precision and recall [4]; fallout, error, accuracy etc.

3. Classifiers

3.1 KNN (K- Nearest neighbor)

K nearest neighbors is an elegant supervised machine learning algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions).K-NN works on a principle that the points (documents) which are close in the space belong to the same class.

The algorithm assimilates all training samples and predicts the response for a new sample by analyzing a certain number (K) of the nearest neighbors of the sample by using some similarity measure such as Euclidean distance measure etc., the distance between two neighbors using Euclidean distance can be found using the given formula.

$$\boldsymbol{Dist(x, y) = \sqrt{\sum_{i=1}^{d}(xi - yi)\text{^}2}}$$

A major demerit of the similarity measure used in k-NN is that it uses all features in computing distances which degrades its performance. In myriad document data sets, only smaller number of the total vocabulary may be useful in categorizing documents [2].

A probable approach to tackle this problem is to learn weights for different features (or words in document data etc.) [9]. Proposed Weight Adjusted k-Nearest Neighbor (WAKNN) classification algorithm is based on the k-NN classification paradigm which can enhance the performance of text classification.

3.2 Support vector machine

Support Vector Machine (SVM), is one of most efficient machine learning algorithm. The original SVM algorithm was invented by Vladimir N. Vapnik and Alexey Ya. Chervonenkis in 1963 and used mostly for pattern recognition. This has also been applied to many pattern classification problems such as image recognition, speech recognition, text categorization, face detection and faulty card detection, etc.

Pattern recognition aims to classify data based on either a priori knowledge or statistical information extracted from raw data, which is a powerful tool in data separation in many disciplines. SVM is a supervised type of machine learning. algorithm in which, given a set of training examples, each marked as belonging to one of the many categories, an SVM training algorithm builds a model that predicts the category of the new example. [3]

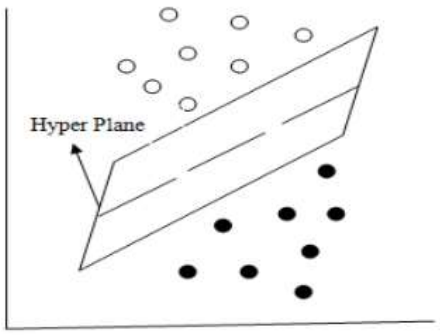

Figure 2: Hyper plane

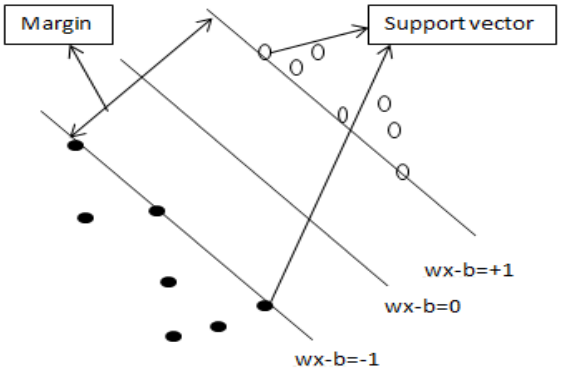

Figure 3: SVM model

The figure 3 is the simple model for representing support vector machine technique. The model consists of two different patterns and the goal of SVM is to separate these two patterns. The model consists of three different lines. [3]

The line w.x-b=0 is known as margin of separation or marginal line.The lines w.x - b = 1 and w.x – b = -1 are the lines on the either side of the line of margin. These three lines together construct the hyper plane that separates the given patterns and the pattern that lies on the edges of the hyper plane is called support vectors.

The perpendicular distance between the line of margin and the edges of hyper plane is known as margin. One of the objectives of SVM for accurate classification is to maximize this margin for better classification. The larger the value of margin or the perpendicular distance, the better is the classification process and hence minimizing the occurrence of error.

3.3 Naïve bayes

The Naïve Bayes classifier is a probabilistic classifier based on Bayes theorem with strong and naïve independence assumptions. It is supposed to be one of the most basic text classification techniques with various applications in email spam detection, personal email sorting, document categorization, language detection and sentiment detection. Experiments describes that this algorithm performs well on numeric and textual data. Though it is often outperformed by other techniques such as boosted trees, random forests, Max Entropy, Support Vector Machines etc., Naïve Bayes classifier is quite efficient since it is less computationally intensive and it necessitates a small amount of training data. The assumption of conditional independence is breached by real-world data with highly correlated features thereby degrading its performance.

3.4 Neural network

Neural networks can be used to model complex relationships between inputs and outputs to find patterns in data. By using neural networks as a tool, data warehousing firms are gathering information from datasets in the process known as data mining. A neural network classifier is a network of units, where the input units usually represent terms, the output unit represents the category. For classifying a text document, its term weights are assigned to the input units; the activation of these units is propagated forward through the network, and the value that the output unit takes up as a consequence determines the categorization decision. [5]

3.5 Decision tree

When decision tree is used for text classification it consist tree internal node are label by term, branches departing from them are labeled by test on the weight, and leaf node are represent corresponding class labels. [10]

Tree can classify the document by running through the query structure from root to until it reaches a certain leaf, which represents the goal for the classification of the document. Most of training data will not fit in memory decision tree construction it becomes inefficient due to swapping of training tuples.

4. Comparative observation

The performance of a classification algorithm is most affected by the quality of data source. Irrelevant and redundant features of data not only increase the cost of mining process, but also reduce the Quality of the result in some cases [8]. Each algorithm has its own advantages and disadvantages as described in Table. The works in [6] compare the most common method in most cases support vector machine and K-nearest neighbor have better accurate result neural network is after then and then naïve bays is last and its evaluation index is again break –even point. Decision tree is less effective then all algorithms describe above.

5. Conclusion

Text classification is very helpful in the field of text mining, The volume of electronic information is increase Day by Day and its extracting knowledge from these large volumes of data. The classification problem is the most essential problems in the machine learning along with data mining literature. This paper survey on text classification. This survey focused on the existing literature and explored the documents representation and an analysis classification algorithms Term weighting is one of the most vital parts for construct a text classifier. The existing classification methods are compared based on pros and cons. From the above discussion it is understood that no single representation scheme and classifier can be mentioned as a general model for any application Different algorithms perform differently depending on data collection.

Table1. Comparison between Classification Algorithms

Classification algorithm	Pros	Cons
KNN	• Effective • Non-parametric	• Classification time is long • Difficult to find optimal k
SVM	• Work well on numeric or textual data • Easy to implement and computation • Work for both linear and nonlinear data. • More capable to solve multi- label classification.	• Perform very poorly [8]when features are highly correlated
Naïve bayes	• Work well on numeric or textual data • Easy to implement and computation Compare with different algorithm	• Perform very poorly when features are highly correlated
Neural network	• Provide better result in complex domain • Testing is very high	• Long training process
[1]Decision tree	• Easy to understand • Easy to generate rule • Reduce problem complexity	• Training time is expensive • A document is only connected with one branch • Once mistake is made at higher level, any sub tree is wrong • May suffer from over fitting

6. References

[1] Basant Agarwal and Namita Mittal, "Text Classification Using Machine Learning Methods-A Survey," *springer*, 2014.

[2] Gongde Guo, Hui Wang, David Bell, Yaxin Bi, Kieran Greer, "KNN Model-Based Approach in Classification," *springer*, 2003.

[3] A. Pradhan, "SUPPORT VECTOR MACHINE-A Survey," *International Journal of Emerging Technology and Advanced Engineering* , vol. 2, no. 8.

[4] Anuradha Patra1, Divakar singh2, "A Survey Report on Text Classification with Different Term Weighing Methods and Comparison between Classification Algorithms," *International Journal of Computer Application*, vol. 75, no. 7, 2013.

[5] Ramasundram, S.P.Victor, "Text Categorization by Backpropagation Network," *International Journal of Computer Applications*, vol. 8, no. 6, 2010.

[6] Mahender, Vandana Korde C Namrata, "TEXTCLASSIFICATIONANDCLASSIFIERS: ASURVEY," *International Journal of Artificial Intelligence & Applications*, vol. 3, no. 2, 2012.

[7] P. D. B. H. A. Dasgupta, "“Feature Selection Methods for Text Classification," *ACM*, 2007.

[8] Dr.K.Prabha S.Brindha Dr.S.Sukumaran, "A SURVEY ON CLASSIFICATION TECHNIQUES FOR TEXT MINING," *IEEE*, 2016.

[9] K. Gayathri ,A.marimuthu, "Text Document Pre-Processing with the KNN for Classification Using the SVM," *IEEE*, 2012.

[10] F. SEBASTIANI, "Machine Learning in Automated Text Categorization," *ACM*, vol. 34, no. 1.