PERFORMANCE ANALYSIS OF TEXT CLASSIFIERS BASED ON NEWS ARTICLES-A SURVEY

SAMPADA BIRADAR[a1] AND M. M. RAIKAR[b]

[ab]School of Computer Science & Engineering, KLE Technological University,Vidyanagar, Hubballi, Karnataka, India

ABSTRACT

In data mining Text classification is one of the important application. Present study classifies news data and detects particular news category. News article classification is a growing interest in the research of text mining. Precisely identifying the news into particular category is a big challenge because of large amount of features in the dataset. In this article classification is one of the important factor. It Classifies news data into four classes namely business, sports, entertainment, and politics. Different classifiers are applied to news data set to classify the news categories and measure the performance. This paper presents a survey of news article classification. There are various Text Classifiers available and all of them vary in efficiency and the speed with which they classify documents. The news articles classification is discussed in this paper using Support Vector Machine classifier, Naïve Bayes classifier, Decision Tree classifier, Neural Network and Random Forest.

KEYWORDS: Data Preprocessing, Text Classification, News Article, Classification Algorithms, Text Mining.

The text mining are gaining more importance recently because of the availability of the increasing number of the electronic documents from a variety of sources which includes unstructured and semi structured information. The main goal of text mining is to enable users to extract information from textual resources and to deal with the various operations like, retrieval, classification (supervised, unsupervised and semi supervised).Text classification is an important part of text mining, it is done on the basis of words, phrases and word combinations with respect to set of predefined class labels. Text Classification processes consists of training phase and testing phase. In training phase, dataset is loaded and different classification algorithms are applied to this dataset. After completing the training phase, performance of classifiers is analyzed and the classifier which provides the best performance is selected [6]. Data mining is useful for extracting or discovering new relation, hidden knowledge and important patterns from huge amount of data. Data mining is also known as Knowledge Discovery in Databases (KDD). Data mining uses different technique for knowledge discovery such as classification, clustering, summarization, associations etc. Text mining is one of the most important technique used in data mining for analysis of large volume of textual data and it is also one of the key technologies used in text mining.

The documents are classified using text classification techniques. This technique is important for categorization of documents in a supervised way[3]. Present research uses text classification technique for classification of news. In this study news data is classified as per the types of news such as business, sports, entertainment and technology.

This technique works in two stages. In first stage, it can extract subsequent terms or effective keywords which are useful for identifying class in training phase. In next stage i.e. testing phase actual classification of document is carried out using subsequent terms of keywords. For effectiveness and efficiency purpose these documents are pre-processed. Text is classified using keyword extraction technique[3][7]. The data is pre-processed by removing stop words which uses stemming technique. After pre-processing frequency for each term in text document is calculated and TF-IDF is found. Experiments have proved that various classifiers can provide higher accuracy.

The paper is organized in the following section: In Section I we have briefly introduced about the text mining, Section II describes the Literature Survey, Section III describes the proposed architecture design of news classification, Section IV discusses the details of classifiers (Neural Network, Random Forest, Naïve Bayes, SVM, Decision Tree). Section V describes the conclusion and last section describes the references used in this paper.

LITERATURE SURVEY

Chy, Abu Nowshed, in this paper they have described about an approach that provides a user to find out news articles which are related to a specific classification. The naive Bayes classifier is used for

[1]Corresponding author

classification of Bangla news article contents based on news code of IPTC. The experimental result shows the effectiveness of classification system [1].

Fabrizio, Sebastiani**,** in this paper the survey discusses about the main approaches to text categorization that fall within the machine learning paradigm. We will discuss in detail issues pertaining to three different problems, namely, document representation, classifier construction, and classifier evaluation. [2].

Menaka S and N. Radha in this paper they have represented using Naive Bayes, Decision tree and K-Nearest Neighbor (KNN) algorithms and its performance are analyzed. Decision tree algorithm gives the better accuracy for text classification when compared to other algorithms. [3].

Aggarwal, Charu C, in this paper has provided a survey of a wide variety of text classification algorithms [4].

Bo pang, Lillian Lee, in this paper they have made conclusion by examining factors that make the sentiment classification problem more challenging [5].

Muhammad Bilal, author has analyzed SMS spam to identify novel features that distinguishes it from SMS (ham). The novelty of their approach is that they intercept the SMS at the access layer of a mobile phone - in hexadecimal format and extract two features: (1) octet bigrams (2) frequency distribution of octets.They evaluate the detection rate and false alarm rate of our system using different classifiers on a real world dataset [6].

Yang, Jian Zhu, in this paper has proposed a new keywords extraction method based on text classification[7].

Dewi Y,This research employs Support Vector Machine (SVM) to classify Indonesian news. SVM is a robust method to classify binary classes. The experiment has proved that SVM provides good performance measure [8].

Kaur, in this paper presents a system for the classification of news articles based on artificial neural networks and have compared the results with the previously used techniques for classification [9].

Wang, Yaguang, et al, in this paper it has been found that Naive Bayes classifier has a higher accuracy and rate by classifying Movie Reviews in NLTK using Decision Tree classifier, Naive Bayes classifier, Maximum Entropy classifier and K-nearest neighbor classifier [10].

A Balahur, in this article presents a comparative study on the methods and resources that can be employed for mining opinions from quotations in newspaper articles.They conclude that a generic opinion mining system requires both the use of large lexicons, as well as specialised training and testing data. [11].

Dilrukshi, in this paper presents a practical experiment to choose a high perform classification method and the theoretical reasons for the high performed classification [12].

Sunita Beniwal, This paper is an introductory paper on different techniques used for classification and feature selection. [13].

Kannan, in this paper the objective of this study is to analyze the issues of preprocessing methods such as Tokenization, Stop word removal and Stemming for the text documents [14].

Gurmeet, in this paper presents algorithm for category idenification of news and have analysed the shortcomings of a number of algorithm approaches[15].

Aamer, in this paper, classifies sentiment analysis of user opinion through comments and tweets using Support Vector Machine (SVM). The goal is to develop a classifier that performs sentiment analysis, by labeling the users comment to positive or negative. From which it is easy to classify text into classes of interest[16].

Jakkula, in this tutorial presents a brief introduction to SVM.[17].

PROPOSED DESIGN

The proposed design represents the news article classification process which is achieved through various steps. These steps involves news data gathering, data pre-processing, feature selection, classification and performance anlysis. (figure 3.1)

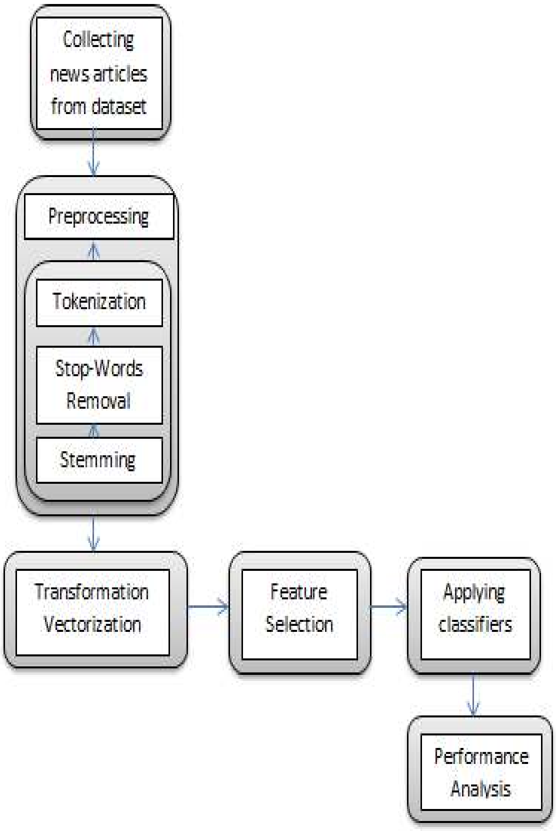

Figure 3.1: System Architecture of News Article Classification

Text Pre-Processing

When data is given as input it is necessary to preprocess the data.Text preprocessing is the process of preparing and cleaning the data of dataset for classification. It helps to reduce the noise in the text, improve the performance of the classifier and speed up the classification process. Preprocessing data has following 3 steps

•Tokenization**:** It is a kind of pre-processing where running text is segmented into words or sentences. Before any real text processing is to be done, text needs to be segmented into linguistic units such as words, punctuation, numbers, alpha-numerics, etc.

This process is called tokenization. Tokenization, when applied to documents, is the process of substituting a sensitive data element with a non-sensitive equivalent, referred as token that has no extrinsic or exploitable meaning or value. A document is considered as a string, and then partitioned into a list of tokens. Stop words such as “the”, “a”, “and”, etc. are frequently occurring; therefore the insignificant words need to be removed.

•Stop word removal: In computing, stop words are words which are filtered out before or after processing of natural language data (text).

•Stop words usually refer to the most common words in a language. The most common words are in text documents are prepositions, articles, and pro-nouns etc, that does not provide the meaning of the documents. These words as treated as stop words. Example for stop words: the, in, a, an, with, etc. [7] Hence it is necessary to remove those words which appear too frequently that provide no information for the task. Stop words are removed in order to save both time and space. Stop words are an integral part of information retrieval process. The removal of stop words increases performance and search results. The stop words needs to be removed for a reason since they provide no distinctive information for classification purpose.

• Stemming: It is the process for reducing derived words to their stem, or root form i.e. it mainly removes various suffixes as a result in the reduction of number of words. For Example, the words user, users, used, using all can be reduced to the word “USE”. This will reduce the required time space [15].

Transformation

It is the extraction of features in a format supported by machine learning algorithms from datasets. Following are the machine learning techniques that has to be applied before the text is sent for classification.

•TF-IDF**:** In information retrieval or text mining, the term frequency–inverse document frequency (also called tf-idf), is a well known method to evaluate how important is a word in a document. Tf-idf is a very interesting way to convert the textual representation of information into sparse features.

The weight of each word is calculated with the help of TF-IDF. TF-IDF calculates values for each word in a document defined as below –

$wd = fw, d* \log(|D| fw,D)$,

w represent words, D is collection of documents, d is individual document belongs to D, |D| is size, fw,d-is number of times w appears in d, fw, -D is number of documents in which w occurs in D[16].

• Count vectorizer**:** It turns a collection of text documents into numerical feature vectors.

•Hash vectorizer**:** In machine learning, feature hashing, also known as the hashing trick , is a fast and space-efficient way of vectorizing features, i.e. turning arbitrary features into indices in a vector or matrix. It works by applying a hash function to the features and using their hash values as indices directly.

Feature Selection

After preprocessing and Transformation the important step of text classification is feature selection. The main idea of feature selection is to select a subset of features from the original document.

Data contains many features, but all the features may not be relevant so the feature selection is used so as to eliminate the irrelevant features from the data without much loss of the information. Feature selection is also known as attributes selection or variable selection[13].

It is performed by keeping the words with highest score accoring to predetermined measure of the importance of the word.

Classification

The documents can be classified by supervised and unsupervised methods. When the class label of each document is known that is supervised, when the class label of document are not kbown that is called unsupervised.

Performance Measure

This is the last step of news text classification. This is experimentally done, rather than analytically. In this step measure the performance. Many measures have been used like precision and recall.

CLASSIFIER ALGORITHMS

There are various classifiers discussed in this paper for classification of News articles as follows:

Support Vector Machine

Support vector machine can be referred to as supervised machine learning algorithm. Important property of SVM is that their ability to learn can be independent of dimensionality of feature space. It can be used for classification and regression problems. The goal of SVM is to design a hyperplane that classifies all training vectors into two classes.

Support Vector Machine Classifier (SVM).There are several advantages of using SVM to train the system. SVM tends to deal with high dimensional data sets [11][5]. SVM creates a hyper plane between data groups. It creates the hyper plane by maximizing the margin as given in figure 4.1. Margin is the distance from the hyper plane to the closest data points.

SVM do not address to the local minimum of the error rate. This caused to increase the accuracy of SVM [17].

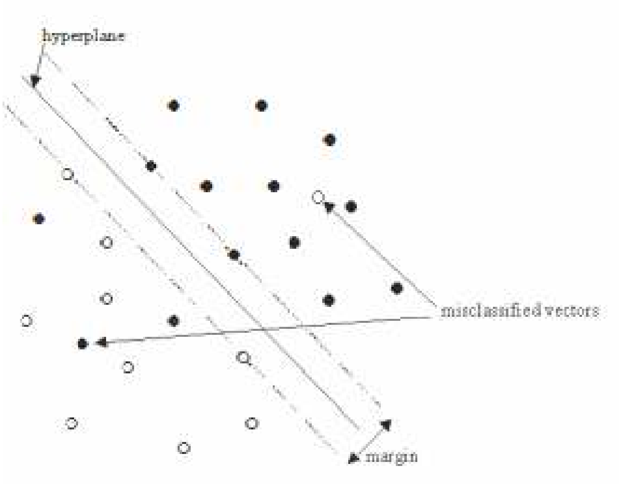

Figure 4.1: Support Vector Machine

Decision Tree Classifiers

Decision tree induction is the learning of decision tree classifiers constructing tree structure where each internal node (no leaf node) denotes attribute test. Each branch represents test outcome and each external node (leaf node) denotes class prediction. At every node, the algorithm selects best partition data attribute to individual classes.

Decision trees are one of the most widely used machine learning algorithms. They are popular because they can be adapted to almost any type of data. They are a supervised machine learning algorithm that divides its training data into smaller and smaller parts in order to identify patterns that can be used for classification.

Decision trees are built using a heuristic called recursive partitioning. In the training phase the algorithm learns what decisions have to be made in order to split the labelled training data into its respective classes. figure 4.2.

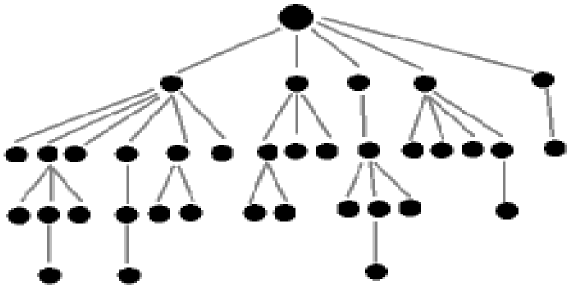

Figure 4.2: Decision Tree

Whenever an unknown label is given, inorder to classify it, the data is passed through the tree. At each decision node a specific feature from the input data is compared with a constant that was identified in the training phase. The decision will be based on whether the feature is greater than or less than the constant, creating a two way split in the tree. The data will eventually pass through these decision nodes until it reaches a leaf node which represents its assigned class.

Neural Network

Neural networks are used in a wide variety of domains for the purpose of classification. The main difference for neural network classifiers is to adapt these classifiers with the use of word features. We note that neural network classifiers are related to SVM classifiers. Each unit receives a set of inputs, which are denoted by the vector Xi, which in this case, correspond to the term frequencies in the ith document. Each neuron is also associated with a set of weights A, which are used in order to compute a function of its inputs. Linear function is as follows: (figure 4.3)

$$pi = A \cdot Xi$$

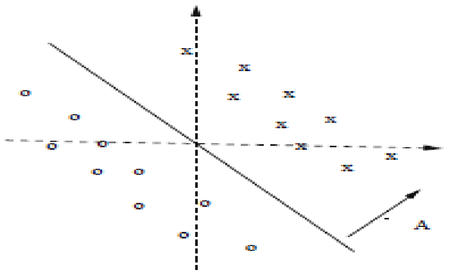

Figure 4.3: Neural Network

Random Forest

Random Forest consists of many classification trees known as tree classifiers, which are used to classifies the news articles based on the categorical dependent on text [18]. Each tree gives a class for the input text documents and the class with highest weight words will be chosen. This classifier' s error rate depends on the correlation between any two trees in the forest and the strength of the each individual tree in the forest. In order to minimize the error rate the trees should be strong and independent of each other [19].

Naïve Bayes Classifier

The naive Bayesian classifier is uncomplicated and widely used method for supervised learning. It is one of the fastest learning algorithms, and can deal with any number of features and classes. Naive Bayesian performs incredibly well in a variety of problems. Furthermore, Naive Bayesian learning is robust enough that small amount of noise does not disturb the results.

CONCLUSION

Text classification is an extensive area in a news article research area. A survey of news article classification is discussed in this paper. The classification steps i.e data gathering, preprocessing, feature selection and classification algorithms are explained. A research survey on various classifiers is done. Moreover these algorithms can be improved and their accuracy of categorization could also be improved in order to achieve better accuracy.

REFERENCES

Nowshed C.A., Seddiqui M.H. and Das S., 2014. "Bangla news classification using naive Bayes classifier." Computer and Information Technology (ICCIT), 2013 16th International Conference on. IEEE, 2014.

Fabrizio S., 2002. "Machine learning in automated text categorization." ACM computing surveys (CSUR), **34**(1):1-47.

Menaka S. and Radha N., 2013. "Text classification using keyword extraction technique." International Journal of Advanced Research in Computer Science and Software Engineering, **3**(12).

Aggarwal C.C. and Cheng X.Z., 2012. "A survey of text classification algorithms." Mining text data. Springer US, pp. 163-222.

Bo P., Lee L. and Vaithyanathan S., 2002. "Thumbs up?: sentiment classification using machine learning techniques." Proceedings of the ACL-

02 conference on Empirical methods in natural language processing, Volume 10. Association for Computational Linguistics.

Unaid M.B. and Farooq M., 2011. "Using evolutionary learning classifiers to do MobileSpam (SMS) filtering." Proceedings of the 13th annual conference on Genetic and evolutionary computation. ACM.

Gong Y.L., Zhu J. and Tang S.P., 2013. "Keywords extraction based on text classification." Advanced Materials Research, Vol. **765**. Trans Tech Publications.

Liliana D.Y., Hardianto A. and Ridok M., 2011. "Indonesian news classification using support vector machine." World Academy of Science, Engineering and Technology, **57**:767-770.

Kaur G. and Bajaj K.., "News Classification using Neural Networks."

Wang Y. et. al., 2015. "Comparison of Four Text Classifiers on Movie Reviews." Applied Computing and Information Technology/2nd International Conference on Computational Science and Intelligence (ACIT-CSI), 3rd International Conference on. IEEE, 2015.

Alexandra B. et. al., 2009. "Opinion mining on newspaper quotations." Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT'09. IEEE/WIC/ACM International Joint Conferences on. Vol. **3**. IEEE, 2009.

Inoshika D. and Zoysa K.D., 2013. "Twitter news classification: Theoretical and practical comparison of SVM against Naive Bayes algorithms." Advances in ICT for Emerging Regions (ICTer), International Conference on. IEEE, 2013.

Beniwal S. and Arora J., 2012. "Classification and feature selection techniques in data mining." International journal of engineering research & technology (ijert) **1**(6).

Kannan S. and Gurusamy V., "Preprocessing Techniques for Text Mining."

Gurmeet K., “News classification and its techniques: A Review”

Khan A.Z.H., Atique M. and Thakare V.M., 2015. “Sentiment Analysis Using Support Vector Machine”, International Journal of Advanced Research in Computer Science and Software Engineering, **5**(4).

Jakkula V., 2006. "Tutorial on support vector machine (svm)." School of EECS, Washington State University, Vol **37**.

Breiman L., 2001. "Random Forests.",Machine Learning, **45**:05-32.

Breiman L. and Cutler A., 2012. "Random Forests." Internet: www.stat.berkeley.edu/-breimanl RandomForests/cc_home.htm.