ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 7, July 2015

Text Mining Approach to Classify Technical Research Documents using Naïve Bayes

Mahesh Kini M[1], Saroja Devi H[2], Prashant G Desai[3], Niranjan Chiplunkar[4]

PG Student, Computer Science and Engineering, NMAMIT, Nitte[1]

Professor and HOD, Department of Computer Science and Engineering, NMAMIT, Nitte[2]

Research Scholar, Department of Computer Science and Engineering, NMAMIT, Nitte[3]

Principal, NMAM Institute of Technology, Nitte[4]

Abstract: World Wide Web is the store house of abundant information available in various electronic forms. Since past few years, the increase in the performance of computers in handling large quantity of text data has led researchers to focus on reliable and optimal retrieval of visible and implied information that exist in the huge resources. In text mining, one of the challenging and growing importance's is given to the task of document classification or text characterization. In this process, reliable text extraction, robust methodologies and efficient algorithms such as Naive Bayes and other made the task of document classification to perform consistently well. Classifying text documents using Bayesian classifiers are among the most successful known algorithms for machine learning. This paper describes implementations of Naïve Bayesian (NB) approach for the automatic classification of Documents restricted to Technical Research documents based on their text contents and its results analysis. We also discuss a comparative analysis of Weighted Bayesian classifier approach with the Naive Bayes classifier.

Keywords: Classification; Naive Bayes; Weighted Bayesian.

I. INTRODUCTION

The development of the internet and the growing number of documents available electronically complicated the management work with large datasets. Automatic classification of text documents or document classification is one such well-known problem in computer science. Further, text document classification (TDC) is an important task in the information retrieval (IR) and natural language processing [9] (NLP) fields and is a process of assigning the predefined categories to text documents. These problems are of great practical importance, given the massive volume of online text available through the World Wide Web, Internet news feeds, electronic mail, corporate databases, medical patient records and digital libraries. The Bayesian Classification presents a supervisory learning [3] method as well as a statistical method of classification, which assumes an underlying probabilistic model that suits very well with text mining. The NB approach, is one of the most effective and straightforward method for text document classification and has exhibited good results in previous studies conducted for data mining.

The task is to assign a document to one or more categories and sub- or subjective categories, based on its text contents. There are two types of classification: supervised and unsupervised. Supervised classification is based on external source, for example, human feedback, which provides information on the correct classification. On the contrary, in un-supervised classification, the processing must be performed entirely without any external information.

This work concentrates only on supervised tasks. A learning model is created to learn the properties or behaviors of documents manually, in some cases semi-automatically, so that more it gets trained through large number of training documents, more efficient is the model to predict or classify the test documents. A schematic text classification model is shown in Fig 1.

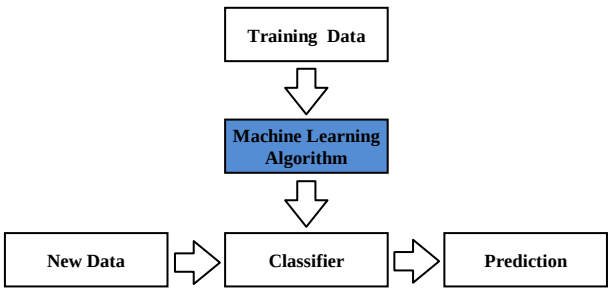

Fig.1 Simplified diagram of the general model building procedure for text classification.

Naïve Bayes method makes use of probability theory to perform the prediction accurately and more efficiently. The disadvantage of Naïve Bayes classifier is that it does not involve morphological relation among the features or terms. This drawback can be overcome by Naïve Bayes variance such as Bernoulli Naïve Bayes, Weighted Naïve Bayes, and NB with semantic probability or ontology analysis.

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 7, July 2015

There is a need to manage various types of documents more effectively as text or document classification. In order to satisfy the need, documents can be divided according to user criteria. This is a reason why it is still essential to be concerned with the classification problem-solving. Methods for document classification have been used extensively over the past two decades but their real importance was acquired just recently. In the past few years, there was a swift development in this area. It is now possible to choose from a various set of classifiers. Further, there are several methods to increase the accuracy of classification. These progresses have made it much easier to deal with classification tasks. However, the arduous task is still in finding a suitable approach for a given problem.

This paper is concerned with the Naive Bayes classifier. Naive Bayes uses a simple probabilistic model that allows inferring the most likely class of an unknown document using Bayes' rule. Because of its simplicity and high accuracy, Naive Bayes is widely used for text classification [4]. It is also considered to be a core technique in information retrieval and we exploit the same in our paper with comparisons to its efficient variance, namely weighted NB.

The implementation problem addressed in this paper is to learn to classify or predict the category/group unlabeled technical text documents. The problem would be solved by taking a large set of labeled (the category/group of the document) technical text documents and building a Naïve Bayes classifier from those documents. The Naïve Bayes classifier would then be able to classify an unlabeled technical document based on the information learned from the labeled document examples.

The rest of the paper is organized as follows. Section II describes text mining and pre-processing phase involved in the text classification and the methods adopted in the implementation. Section III gives the design of our Naive Bayesian Classifier details. Section IV explains implementation of classifier and explains in detail the various phases involved. Section V gives performance evaluation and analysis of classifier and compares with weighted NB. Section VI gives conclusion and pointer to future work.

II. TEXT MINING

Text mining [5] is the process of computation that involves extraction of information from bulk quantity of data and uncovering new unidentified information by retrieving from numerous written and digital resources with the help of algorithms of robust form. A schematic is shown below Fig 2.

There are different algorithms [8] of text mining is available for proficient classification and categorization. Some of them are k-nearest neighbour, Support vector machine and Bayesian classifier and K-mean clustering.

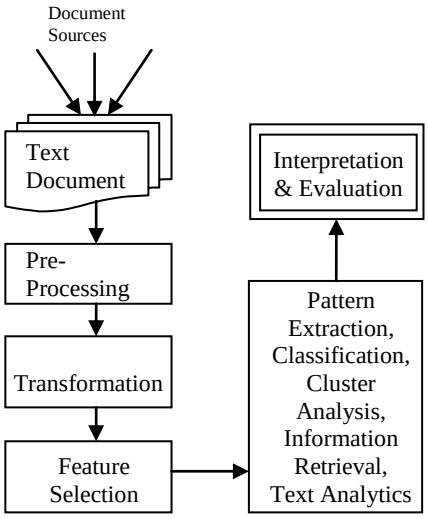

Fig.2. Text mining process flow

A. Text Pre-processing

The text pre-processing step is divided into number of sub tasks as follows:

i) Tokenization: Text document, collection of sentences divided into words by removing spaces, commas or any other delimiters.

ii) Stop word Removal: This step involves removal of common words like 'a', 'of' or any other tags in the collection of tokens.

iii) Stemming: This technique is used to find the root or stem of a word. Stemming converts words to their root words. For example the words like ran, running converted to run [6].

iv) Text Transformation or Feature Generation:
Converting text document into vector space termed as text transformation which can be used for further analysis task effectively.

v) Feature Selection/Attribute Selection:
This phase mainly performs removing features that are considered irrelevant for mining purpose. This procedure give advantage of smaller dataset size, less computations and minimum search space required.

vi) Text mining methods:
Number of text mining methods in data mining had been proposed such as: Classification, Clustering [12], Information retrieval, Topic discovery, Summarization [7], Topic extraction.

vii) Interpretation or Evaluation:
This phase includes Evaluation and Interpretation of results in terms of calculating Precision and Recall, Accuracy, F measure etc.

B. Text Transformation or Feature Generation

Converting text document into vector space termed as text transformation which can be used for further analysis task effectively.

C. Feature Selection/Attribute Selection

This phase mainly performs removing features that are considered irrelevant for mining purpose. This procedure give advantage of smaller dataset size, less computations

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 7, July 2015

and minimum search space required.

D. Text mining methods

Number of text mining methods in data mining had been proposed such as: Classification, Clustering, Information retrieval, Topic discovery, Summarization, Topic extraction.

E. Interpretation or Evaluation

This phase includes Evaluation and Interpretation of results in terms of calculating Precision and Recall, Accuracy, F measure etc.

III. NAIVE BAYESIAN 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. Although simple in model, Naive Bayesian performs incredibly well in a variety of problems. Furthermore, Naive Bayesian learning is robust enough that small amount of noise does not perturb the results.

An abstract schematic diagram of Naïve Bayes (NB) approach for the automatic classification of technical research documents is shown in Fig 3.

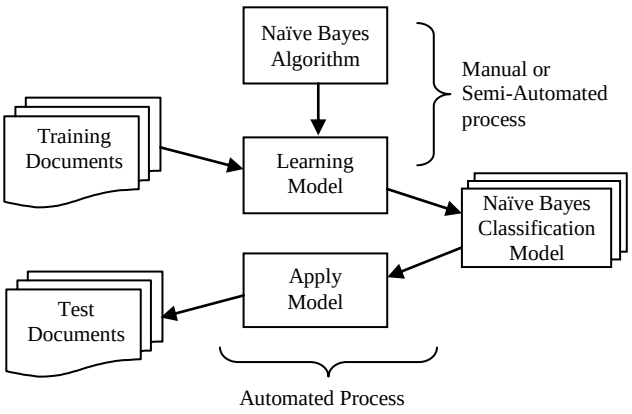

Fig 3. Naïve Bayes Classifier Schematic.

Learning model acquire knowledge in the form of vocabulary for each training document and assign class labels upon which classification model is built. Learning can continue for large number of training documents to make the model more knowledgeable or educated. This model can be then applied on test documents to check for accuracy. Feature extraction is done manually while in learning process where as in testing it is done automatically.

The Naïve Bayesian model is a probabilistic approach to classification which is based on the simplifying assumption of conditional independence among attributes. Given a training set containing attribute values and corresponding target values (classes), the naïve Bayesian classifier predicts the class of an unseen (new) instance, based on previously observed probabilities of the feature terms occurring in that instance.

IV. METHODOLOGY

A. Training the Classifier

We pursue the random sampling for main categories (Technical and Non-Technical) and K fold strategy with k=6 for sub-categories to decide the number of training and testing examples. 92 examples were used as the training set to build the classifier and 46 examples were used to test the classifier for accuracy. The prior probability for each main category (Technical and Non-Technical) is 1/2 (as there are 2 categories) and sub-category is 1/6 (as there are 6 sub-categories). Extraction or selection of the training document is done manually. The posterior probability P(wk|c) is calculated as follows. All documents that belong to respective categories were parsed and a hash table was prepared for each category. All words in the vocabulary served as keys of the hash table. The numeric values of the hash table were the word occurrence frequency (nk) in all documents belonging to that category. The total word count (including repeats) for each category termed as n was also calculated. Posterior probability with Laplace's correction can be calculated using the formula

$$P(wk|c) = (nk +1) / (n + |Vocabulary|).$$

Following are the steps involved in Training the Classifier

1. Let V be the vocabulary of all words in the documents D
2. For each category c_i in C
Let Di be the subset of documents in D in category c_i
$P(ci) = |D_i| / |D|$
Let Ti be the concatenation of all the documents in D_i
Let n_i be the total number of word occurrences in T_i
For each word w_j in V
Let n_{ij} be the number of occurrences of w_j in T_i
Let $P(w_j | c_i) = (n_{ij} + 1) / (n_i + |V|)$

B. Testing the Classifier

To classify a document, for example, D, the probabilities of a given category are look up in the hash table for the words found in D and multiplied together. The category that produces the highest or maximum probability is the classification for document D. Also if a word in D is not present in the original vocabulary (built from training set) the word is ignored. The equation used to classify D is

$$C = \arg\max (P(c) \Pi P (w_k|c)).$$

The Naïve Bayes algorithms to train and test the classifier are as given below:

1. Given a test document D
2. Let n be the number of word occurrences in D
3. Return the category:

$$\underset{ci \text{ in } C}{\text{Arg max}}\ P(c_i) \prod_{i=1}^{n} P(a_i | c)$$

where a_i is the word occurring at the ith position in D.

DOI 10.17148/IJARCCE.2015.4789

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 7, July 2015

Initially the classification model classifies the document as Technical or Non-technical and then further classify to a sub-category automatically to Computer Science, Electronics, Electrical, Mechanical or Civil if technical, or otherwise to Unknown if non-technical. Classification is based on the priori and conditional probability which uses the training set knowledge. Also we observe that Gaussian constant is used while to avoid zero probability for those terms that was not occurred in the training set. Class is assigned to the document for which maximum probability value is calculated.

C. Comparison with Weighted NB Classifier

In order to improve the classification effect of NBC, we propose a method of assigning weight to features. This approach of feature weighting method involves assigning a weight to each feature in naive Bayesian model. In the statistical vector-space model, a document is theoretically represented by a vector of words or texts mined from the document, coupled with weights representing the magnitude of the terms in the document and within the whole document group. The formula of weighted NBC as follow:

$$\text{Arg max}_{ci \text{ in } C} P(c_i) = \prod_{i=1}^{n} wt_i \, P(a_i \mid c)$$

where wt_i is the word weight calculated based the frequency of occurrences in the document.
In text classification, a document may to a certain extent match multiple categories. Best matching category for the text document needed to be found. The term (word) frequency or inverse document frequency (TF-IDF) approach is generally used to weight each word in the text document which finds relevancies among words or terms, text documents and particular categories or sub-categories. Feature weight contributes and influences more as compared to the basic Bayesian classifier in predicting the class labels or categories.

The result is that weight will improve the classification outcome. That is, it strengthens the attributes, which have elevated relationship with classification and weakens attributes that have little relationship with classification. And thereby we can have Weighted Naive Bayesian classifier which makes effective classification and improve its overall classification learning.

V. EXPERIMENTAL RESULTS AND ANALYSYS

A. Test results of a document

The text classifier specified in the above section has been implemented using Naïve Base method. Whenever the input text is given as a pdf or txt file, it will be automatically categorizes the document based on the training set and learning model. We consider a sample documents to discuss the working mechanism of the said classifier as shown in Fig. 4 and it is given as a pdf files.

Initially model classifies the document under test to main category as technical or non-technical one, then it classified to one of the sub-categories. Hence we have a hierarchical class label from Technical to Computer, Electronics, Electrical, Mechanical or Civil and from Non-technical to Unknown. This hierarchical text classification model can be extended to any number of categories and sub-categories to make it more flexible and reliable classification.

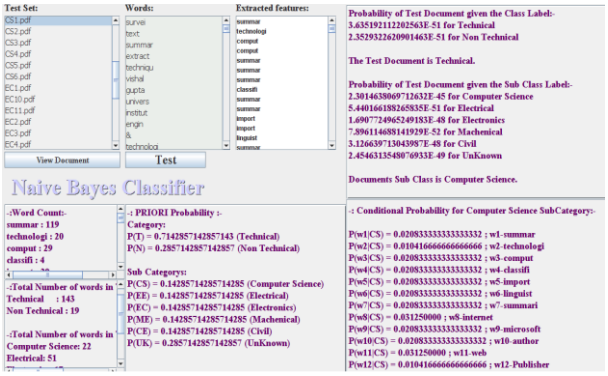

Fig 4. Test results on Document.

B. Classifier Model Evaluation

To test and evaluate the model, 60% of the dataset are used. Instances are extracted and then served as a benchmarking dataset for machine learning problems. By comparing the actual class of the instance with the predicted one (i.e. generated by the classifier model), system performance can be measures in term of recall, precision, and F-measure. These can be mathematically defined as below subsections.

Category	Sub-Category	# train docs	# test docs	Total # docs
Technical	Computer	20	10	30
Technical	Electronics	16	8	24
Technical	Civil	16	7	23
Technical	Electrical	14	5	19
Technical	Mechanical	12	6	18
Non Tech	Unknown	14	5	19
Total:		92	46	138

Table 1. Dataset input in this Study.

In order to further evaluate the performance of the proposed pre-processing stage, the results of not pre-process and pre-process are compared. However, if the results are worse than that when no pre-processing phase is conducted (i.e. the classification model is not good enough), therefore adjusting and fine-tuning parameters are required (e.g. modifying the technique used in feature selection) and hence re-build the model again. This step will stop until a good classification result is obtained. Furthermore, Naïve Bayes classifier will be tested with other classifier such as Decision tree to determine whether Naïve Bayes is the best classifier among them.

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 7, July 2015

B. Performance Evaluation

I. Data Description

The objective of this study is to classify the given specified test documents into six categories, namely, Computer, Civil, Mechanical, Electrical, Electronics, and Unknown correctly. To start with, it is given 10 documents for each category to serve as the dataset for generating the classification model. To build and evaluate the classification model, the total 50 documents will be split into two datasets, namely training set and testing set, in which 65% of the documents will go to the training set whereas the remaining 35% will go to the testing set. In the representation of these documents, they have been vectorized into 8500 terms or word features. No missing data is among the attributes and all the text attributes are described in the term frequency. An example of the data is presented in Table 2 which sum up the description of data in the both training and testing set.

	Training Docs	Testing Docs
# instances	92	46
# terms/words	8500	3400
Missing Data	Nil	Nil

Table 2. Data Description in this Study.

II. Results and Discussions

The objective of this evaluation is twofold. First, it determines whether the pre-processing phase is useful to deduce better classification accuracy and performance when compared to the situation that has not been pre-processed the data. Second, it compares the classification accuracy and performance when different classifiers are applied.

	Correctly Classified Instances	Incorrect Classified Instances	Precision	Recall	F-measure
With Pre-processing and feature selection	96.9%	3.1%	0.969	0.969	0.969
Without Pre-processing & feature selection	95.5%	4.5%	0.956	0.955	0.955

Table 3. Classification Accuracy of Naïve Bayes Classifier

Nearly 65% data (i.e. 46 documents) are extracted randomly to build the training dataset for the classifier. A dataset with 138 documents classified in two main and six different sub-categories is used for evaluation. The other 46 documents are used as the testing dataset to test the classifier. Table 3 summarizes the result of using Naïve Bayes classifier to classify the documents.

VI. CONCLUSION AND FUTURE WORK

Classification is a very common and extensive task in information processing, and it is not generally easy. Application areas of text or document classification are email spam, opinion mining, labelling, text object recognition from news document. While these tasks might be easy for human beings they are very hard for machines. Even with the simplest cases there are noise and distortions distressing the measurement results and making the classification task nontrivial.

A further step is to perform a syntactic analysis and tag each word with its part of speech to involve morphological relationships among the features. This helps to disambiguate different senses of a word and to eliminate incorrect analyses caused by rare word senses. Also a Concept-Based Analysis [10] among documents can improve the overall classification effect. This measure quantifies to the concepts that only appear in a small number of documents as these concepts can discriminate their documents among others. Text Classification of documents is of robust nature can also be applied to automatic question-answering system [11].

It is encouraging to observe that Naive Bayes approach used in this paper enhances the richness of features of a technical research document for classification. It categorizes the document into very broad categories. NB approach for classification of technical research document for the six sub-categories considered above yielded 88.05% accuracy. It is also observed that the classification accuracy of the classifier is relative to number of training documents.

The results are reasonably encouraging. This approach can be used by search engines for useful categorization of websites to build an automated website directory based on type of organization. However in this experiment, distinct and hierarchical categories are considered. The same algorithm could also be improved to classify the pages into more specific categories (hierarchical classification) by changing the feature set e.g. a web site that is academic may be further classified into school, college or a university website.

REFERENCES

[1] Ajay S. Patil, B.V. Pawar, "Automated Classification of Web Sites using Naive Bayesian Algorithm", Proceedings of the International MultiConference of Engineers and Computer Scientists, Hong Kong 2012, Vol I, 14-16.

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 7, July 2015

[2] DUAN Wei1, LU Xiang, "Weighted naive Bayesian classifier model based on information gain", International Conference on Intelligent System Design and Engineering Application, 2010, pp. 819-822.
[3] Falguni N. Patel, Neha R. Soni, "Text mining: A Brief survey", International Journal of Advanced Computer Research (ISSN) Volume-2 Number-4 Issue-6,243-248, Dec-2012.
[4] Vishal Gupta and Guruprit Lehal, "A Survey of Text Mining Techniques and Applications", Journal Of Emerging Technologies In Web Intelligence, Vol. 1, No. 1, 60-76, August 2009.
[5] N. Kanya and S. Geetha, "Information Extraction: A Text Mining Approach", IET-UK International Conference on Information and Comm. Technology, IEEE, Dr. M.G.R. University, Chennai, India, 1111- 1118, 2007.
[6] M. Porter, "An algorithm for suffix Stripping and stemming", A Text Book, pages 130–137, 1980.
[7] C. Lakshmi Devasenal and M. Hemalatha, "Automatic Text Categorization and Summarization using Rule Reduction", IEEE-International Conference on Advances In Engineering Science & Management,594-598, March 2012.
[8] Mr. Rahul Patel , Mr. Gaurav Sharma, "A survey on text mining techniques", International Journal Of Engineering And Computer Science,Volume 3 Issue 5, 5621-5625, May-2014.
[9] Shaidah Jusoh and Hejab M. Alfawareh, "Techniques, Applications and Challenging Issue in Text Mining", IJCSI International Journal of Computer Science Issues, Saudi Arabia Vol. 9, Issue 6, No 2, Nov-2012
[10] Shady Shehata, Member, Fakhri Karray, Senior Member and Mohamed S. Kamel, ―An Efficient Concept-Based Mining Model for Enhancing Text Clustering‖, IEEE Transactions On Knowledge And Data Engineering, Vol. 22, No. 10, Oct-2010.
[11] PayaJ Biswas, Aditi Sharan, Nidhi Malik, "A Framework for restricted domain question answering system", IEEE- International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), 613-620, 2014.
[12] Liritano S. and Ruffolo M., "Managing the Knowledge Contained in Electronic Documents: a Clustering Method for Text Mining", IEEE 454-458, Italy, 2001.

DOI 10.17148/IJARCCE.2015.4789