ISSN: 2454-4248
142 – 146

An Overview on Implementation Using Hybrid Naïve Bayes Algorithm for Text Categorization

Kshitija Deshmukh
Department of Computer Science and Engineering,
Datta Meghe Institute of Engineering and Technology,
Wardha,India
rockjuhi8@gmail.com

Shweta Raut
Department of Computer Science and Engineering,
Datta Meghe Institute of Engineering and Technology,
Wardha,India
shwetaraut567@gmail.com

Jaya Bhargaw
Department of Computer Science and Engineering,
Datta Meghe Institute of Engineering and Technology,
Wardha,India
jayabhargaw02@gmail.com

Asst.Prof. Anand Saurkar
Prof. in Department of Computer Science and Engineering,
Datta Meghe Institute of Engineering and Technology,
Wardha,India
saurkaranand@gmail.com

Abstract:- Automated Text categorization and class prediction is important for text categorization to reduce the feature size and to speed up the learning process of classifiers .Text classification is a growing interest in the research of text mining. Correctly identifying the Text into particular category is still presenting challenge because of large and vast amount of features in the dataset. In regards to the present classifying approaches, Naïve Bayes is probably smart at serving as a document classification model thanks to its simplicity. The aim of this Project is to spotlight the performance of Text categorization and sophistication prediction Naïve Bayes in Text classification.

Keywords—*Classification, Mining, Naive Bayes, Dataset. Spotlight, Sophistication*

******

I . INTRODUCTION

Text classification is the problem of automatically assigning zero, one or more of a predefined set of labels to a given segment of free text. The labels are to be chosen to reflect the "meaning" of the text. Selecting the appropriate set of labels may be ambiguous even for a human rater. When a machine is to try and mimic the human behavior, the algorithm will have to cope with a large amount of uncertainty coming from various sources. First of all, on a purely lexicographic level, human language is ambiguous, including words and word combinations with multiple senses which are disambiguated by the context. More importantly, the definition of meaning of a text is still vaguely defined, and a matter of debate. One does not want to answer the question whether a computer has "understood" a text, but rather – operationally – whether it can provide a result which is comparable to what a human would provide (and find useful) [10].

With the increasing availability of text documents in electronic form, it is of great importance to label the contents with a predefined set of thematic categories in an automatic way, what is also called as automated Text Categorization. In last decades, a growing number of advanced machine learning algorithms have been developed to address this challenging task by formulating it as a classification problem. Commonly, an automatic text classifier is built with a learning process from a set of pre labeled documents. Documents need to be represented in a way that is suitable for a general learning process. The most widely used representation is "the bag of words": a document is represented by a vector of features, each of which corresponds to a term or a phrase in a vocabulary collected from a particular data set .The value of each feature element represents the importance of the term in the document, according to a specific feature measurement. A big challenge in text categorization is the learning from high dimensional data. On one hand, tens and hundreds of thousands terms in a document may lead to a high computational burden for the learning process. On the other hand, some irrelevant and redundant features may hurt predictive performance of classifiers for text categorization. To avoid the problem of the "curse of dimensionality" and to hurry up the educational method, it's necessary to perform feature reduction to cut back feature size

ISSN: 2454-4248
142 – 146

A common feature reduction approach for text categorization is feature selection that this paper concentrates on, where only a subset of original features are selected as input to the learning algorithms. In last decades, a number of feature selection methods have been proposed, which can be usually categorized into the following two types of approach: the filter approach and the wrapper approach.[11] The filter approach selects feature subsets based on the general characteristics of the data without involving the learning algorithms that will use the selected features. A score indicating the "importance" of the term is assigned to each individual feature based on an independent evaluation criterion, such as distance measure, entropy measure, dependency measure and consistency measure. Hence, the filter approach only selects a number of top ranked features and ignores the rest. Alternatively, the wrapper approach covetously searches for higher options with associate analysis criterion supported an equivalent learning algorithmic rule. though it's been shown that the wrapper approach typically performs higher than the filter approach, it's far more process price than the filter approach, that generally makes it impractical.

II. REVIEW of LITERATURE

A novel and efficient feature selection framework based on the Information Theory, which aims to rank the features with their discriminative capacity for classification. Author first revisit two information measures: Kullback Leibler divergence and Jeffrey's divergence for binary hypothesis testing, and analyze their asymptotic properties relating to type I and type II errors of a Bayesian classifier. Author then introduce a new divergence measure, called Jeffery's-Multi-Hypothesis (JMH) divergence, to measure multi-distribution divergence for multi-class classification. Based on the JMH-divergence [1].

Classification is a data mining technique used to predict group membership for data instances within a given dataset. It is used for classifying data into different classes by considering some constrains. The problem of data classification has many applications in various fields of data mining. This is because the problem aims at learning the relationship betAuthoren a set of feature variables and a target variable of interest. Classification is considered as an example of supervised learning as training data associated with class labels is given as input. This paper focuses on study of various classification techniques, their advantages and disadvantages [2].

Approach uses Frequency Ratio Accumulation Method (FRAM) as a classifier. Its features are represented using bag of word technique and an improved Term Frequency (TF) technique is used in features selection. The proposed approach is tested with known datasets. The experiments are done without both of normalization and stemming, with one of them, and with both of them. The obtained results of proposed approach are generally improved compared to existing techniques.The performance attributes of proposed Arabic Text Categorization approach Authorre considered: Accuracy, Recall, Precision and F-measure [3].

The text categorization method to predict the trend of the stock. Author divide the text categorization method into the following three steps: Text representation, Feature selection and Text Categorization. By comparing several categorization methods including feature selections and feature spaces, etc., the results show that the SVM method with Information Gain and 1000 feature spaces can get the better performance for the predict of the stock with the news[4].

How to classify and organize the Text based categorization is made use of for document classification with pattern recognition and machine learning. Advantages of a number of classification algorithms have been studied in this paper to classify documents. An example of these algorithms is: Naive Baye's algorithm, K-Nearest Neighbor, Decision Tree etc. This paper presents a comparative study of advantages and disadvantages of the above mentioned classification algorithm

III. OBJECTIVES

Number of applications, it becomes necessary to improve the efficiency of text categorization algorithm to get much efficient and reliable result.

To do so there is a need to develop or modify existing algorithm or fuse functionality of more than one algorithm.

A] Classification of unknown text in classes.
B] Until now manual classification is done in system and automated classification doesn't give better efficiency.
C] To improve the efficiency of automated text categorization, we can propose a modified approach of Naive Bayes algorithm which outcomes the disadvantages of existing system.

V. PROPOSED SYSTEM

The main objectives of the project are listed below:

1. To preprocess the dataset.
2. To preprocess the newsgroup dataset
3. Generate Frequencies based on Hybrid algorithm
4. Classification of unknown input into a proper group.

IV.PROBLEM STATEMENT

Text categorization plays an important role in many classification systems like disease classification based on symptoms or it may be bug classification. Due to such large

ISSN: 2454-4248
142 – 146

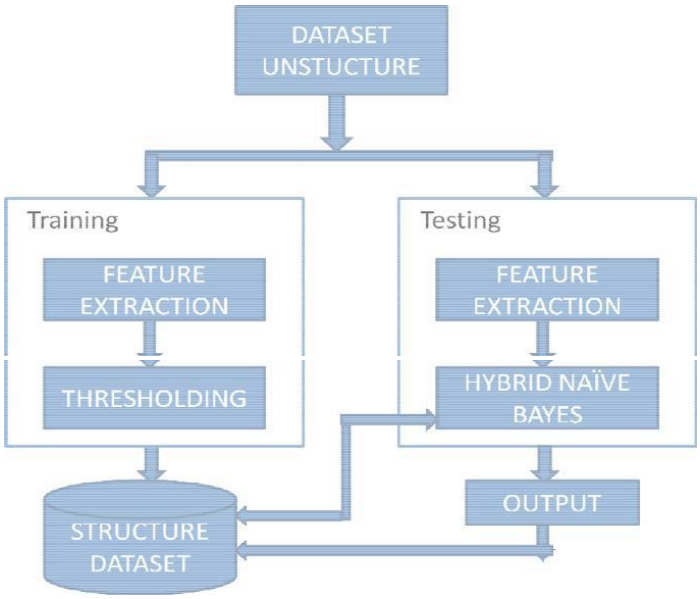

Figure: Architecture Block Diagram

In the architecture, first gives the input file from the given dataset and pass to the training set. In the training set perform the feature extraction that means to remove the stopwords, blank spaces and count total no. of words in total file and also count the if, idf. after that threshold are apply, thresholding are used when only if you required only small data out of large data then we set the threshold percentage when you required then store this data in database.

In testing set, first gives the input file from data and perform same tasks in feature extraction perform in training set except thresholding. Then apply the Hybrid Naïve Bayes algorithm. In this algorithm we calculate the tf, idf, and ntf for accuracy.

Training and the testing module gives the option to user to make selection between the training data and classifying it. Set training data can be used to set any one file into the dataset after pre-processing TF, IDF calculation and other means of processing and frequency generation. All over data is proceeding sequentially and classify the words from dataset with respect to document id, term frequency and inverse document frequency.Classify data can be used to test whether the algorithm is working properly ornnot.This thresholding analysis the data with respect to maximum frequency. It gives the maximum value of frequency on preprocess dataset. Shows the total word from preprocess dataset.

VI. METHODOLOGY

A . Training and Testing Module:

In machine learning, the study and construction of algorithms that may learn from and create predictions on knowledge may be a common task. Such algorithms work by creating data-driven predictions or selections,through building a mathematical model from input file.The data wont to build the ultimate model typically comes from multiple datasets. specifically, 3 knowledge sets are normally utilized in completely different stages of the creation of the model.The model is at first match on a coaching dataset,that is a collection of examples wont to match the parameters of the model. The model (e.g. a neural internet or a naive Thomas Bayes classifier) is trained on the coaching dataset employing a supervised learning technique. In observe,the coaching dataset usually incorporates pairs of Associate in Nursing input vector and also the corresponding answer vector or scalar, that is often denoted as the target. the present model is run with the coaching dataset and produces a result, that is then compared with the target, for every input vector within the coaching dataset. supported the results of the comparison and also the specific learning rule being employed, the parameters of the model are adjusted. The model fitting will embody each variable choice and parameter estimation. Finally, the check dataset may be a dataset accustomed offer Associate in Nursing unbiased analysis of a final model work on the coaching dataset. A check dataset may be a dataset that's freelance ofthe coaching dataset, however thatfollows constant c hance distribution because the coaching dataset. A test dataset is a dataset that is independent of the training dataset,but that follows the same probability distribution as the training dataset. If a model fit to the training dataset also fits the test dataset well, minimal over fitting has taken place . A better fitting of the training dataset as opposed to the test dataset usually points to over fitting.

This holds the number of documents that will be used to model the classifier. Each document from training data will be processed to calculate the term frequency and inverse document frequency. This holds the no of document that will be used as input to test the classifier model which will be classified depending upon the training data.

A.Data Preprocessing

Data Preprocessing means converting unstructured data into structured data. Given a textual source containing different types of documents (different formats, language formatting) the first action that should text preprocessing.[9]

Pre-processor processes the document words by removing:

A] Symbols removal
B] Stop words removal

The all symbols are removed in preprocessing step and a stop list is a list of commonly repeated features which appears in

ISSN: 2454-4248
142 – 146

every text document. The common features such as it, he, she and conjunctions such as and, or, but etc. are to be removed because they do not have effect on the categorization process. Stemming is the process of removing affixes (prefixes and suffixes) from the features. It improves the performance of the classifier when the different features are stemmed into a single feature. For example: (convert, converts, converted, and converting) stemming removes different suffixes(s,-ed,ing) to get a single feature. [12]

B. Generating Frequencies

Gnerate dataset frequencies provided below:

Word count

Term Frequency

Normalized term Frequency
Inverse Document Frequency

Word Count: A word count is a numerical count of how many words a document contains. Most word processors today can count how many words are in a document for the user. Word counting may be needed when a text is required to stay within certain numbers of words. This may particularly be the case in academia, legal proceedings, journalism and advertising.

Term Frequency(TF) : Term Frequency which measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the document length (aka. the total number of terms in the document) as a way of normalization [9].

Normalized Term Frequency(NTF):

Normalized Term Frequency**,** which measures how important a term is. While computing TF, all terms are considered equally important. However it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing the following [9].

Inverse Document Frequency(IDF): The inverse document frequency is a measure of how much information the word provides, that is, whether the term is common or rare across all documents. It is the logarithmically scaled inverse fraction of the documents that contain the word, obtained by dividing the total number of documents by the number of documents containing the term,and taking logarithm of that quotient [9].

C. Classificatio
n a. Naive
Bayes

Naive Bayes Classifiers are simple probabilistic classifiers based on the Bayes Theorem. These are highly scalable classifiers involves a family of algorithms based on a common principle assuming that the value of a particular feature is independent of the value of any other feature, given the class variable. In practice, the independence assumption is often violated, but Naive Bayes classifiers still tend to perform very well under this unrealistic assumption and very popular till date.

In machine learning, naive Bayes classifiers are a family of

simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. The Bayesian Classification

represents a supervised learning method as well as a statistical method for classification. Assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems. The Bayesian classification is used as a probabilistic learning method (Naive Bayes text classification). Naive Bayes classifiers are among the most successful known algorithms for learning to classify text documents. [13]

VII .HYBRID NAIVE BAYES(HNB)

The Hybrid classifier tries to capture the desirable

properties of the Naive Bayes classifiers.

In machine learning, Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

TF*IDF is the product of two statistics, term frequency and inverse document frequency. Various ways for determining the exact values of both statistics exist. Variations of the tf*idf weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query.tf*idf can be successfully used for stop words ,filtering in various subject fields including text summarization and classification. One of the simplest ranking functions computed by summing the tf*idf for each query term; many more sophisticated ranking functions are variants of this simple model.Naive Bayes drawback is give the minimum accuracy and

ISSN: 2454-4248
142 – 146

more time complexity. For this reason ,the naive Bayes usually refers to the HNB classifier

Hybrid Naïve Bayes algorithm calculate:

The inverse document frequency is a measure of how much information the word provides, that is, whether the term is common or rare across all documents and Normalized Term Frequency**,** which measures how important a term is so it give the better accuracy and less time complexity than Naïve Bayes Algorithm. This structure uses the database, test data as input and training data and calculates the probability for each document to go in each group. The one with maximum probability will be the output. In the context of text classification, features or attributes usually mean significant words, multi-words or frequently occurring words indicative of the text category. After feature selection, the text document is represented as a document vector, and an appropriate machine learning algorithm is used to train the text classifier.

CONCLUSION

This application automates the text classification process otherwise would take long time doing manually the same task. Text file are appropriately classified using this application. This application allows you to select the test data, training data. A similar concept can be used for different purposes like arrange your computer, classify various documents with various applications and analyse them. Using Modified Naive Bayes Algorithm is used so that more accuracy and less time complexity can be achieved than that of Naïve Bayes algorithm. System will be provided automatic text categorization using Hybrid Naive Bayes algorithm.

References

[1] Bo Tang, Student Member, IEEE, Steven Kay, Fellow, IEEE, And Haibo He, Senior Member, IEEE, “Toward Optimal Feature Selection in Naive Bayes for Text Categorization” in IEEE Transactions on Knowledge and Data Engineering, 9 Feb 2016.

[2] Jianxiao Liu, ZonglinTian, Panbiao Liu, JiaAuthori Jiang, “Comparative Study of Classification Algorithm for Text Based Categorization” in IJRET, Volume: 05 Issue: 02 , Feb-2016.

[3] Senthil Kumar B, BhavithaVarma E, “A Survey on Text Categorization” In IJARCCE, Vol. 5, Issue 8, August 2016.

[4] PoojaBolaj, SharvariGovilkar, “A Survey on Text Categorization Techniques for Indian Regional Languages”, in International Journal of Computer Science and Information Technologies, Vol. 7 (2), 2016.

[5] Shaifali Gupta, Reena Rani, “Improvement in KNN Classifier (imp-KNN) for Text Categorization”, in IJARCSSE, Volume 6, Issue 6, June 2016.

[6] Sayali D. Jadhav, H. P. Channe “Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques” in International Journal of Science and Research, Volume 5 Issue 1, January 2016.

[7] Rouhia M. Sallam, “Improving Arabic Text Categorization using Normalization and Stemming Techniques” in International Journal of Computer Applications, Volume 135 – No.2, February 2016.

[8] Bozhao Li, Na Chen, Jing Authorn, Xuebo Jin and Yan Shi, “Text Categorization System for Stock Prediction” in IJUNESST, Vol.8, No.2 (2015).

[9] “Indian Language Text Representation and Categorization using Supervised Learning Algorithm” in the Proceedings of the International Conference on Intelligent Computing Applications (ICICA-14), organized by Bharathiar University,Coimbatore, Tamilnadu, during 6th and 7th March 2014. This attached with IEEE Affiliation published by IEEE -CPS and the corresponding ISBN 978-1-47993966-4.

[10] Toni Giorgino , “An Introduction to Text Classication” handbook 1 December 2004.

[11] H. Liu and L. Yu, “Toward integrating feature selection algorithms for classification and clustering,” IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 4, pp. 491–502, 2005

[12] Swati , Meenakshi ,” SVM BASED IMPROVEMENT IN KNN FOR TEXT CATEGORIZATION”, International Journal of Engineering Research and General Science Volume 3, Issue 4, July-August, 2015

[13] Likhita Mate, Priyanka Dudhe,” Efficient Text Categorization using Naïve Bayes Classification”, International Journal of Innovative Research in Science, Engineering and Technology Vol. 6, Special Issue 11, May 2017