Comparison of Text Classification Models for Telugu News Articles

Naga Sudha D[1], Y Madhavee Latha [2]
[1]*Research Scholar JNTUH, Hyderabad*
[2]*JNTU Hyderabad*

***Abstract*- Text classification is become important when the information is increasing rapidly over the internet. This information is in unstructured form and need to be digitized. As these documents are digital form it is necessary for organizing the data by automatically assigning a set of documents into predefined labels based on their content. It mainly depends on the methods that should be used in each phase improves the efficiency of the document classification. In this paper we propose a classification model that supports both the generality and efficiency. It also discusses some of the major issues involved in automatic text classification such as dealing with unstructured text, handling large number of attributes and natural language processing based techniques, dealing with missing metadata and choice of a suitable machine learning technique for training a text classifier. Both are achieved by following the logical sequence of the process of classifying the unstructured text document step by step and efficiency through various methods are proposed. The experimental results over news articles have been validated using statistical measures of accuracy and F-Score. The results have proven that the methods significantly improve the performance.**

***Index Terms*- Text classification, Logistic regression, Naive Bayes classifier, Support Vector machine, Gradient descent trees.**

INTRODUCTION

As the information is increasing exponentially it is necessary to analyze and classify these volumes of data. This makes the importance of text classification begins to spring up. Text classification is the process of assigning the labels to the documents based on their content by building a model through a training data. It also considers the set of predefined labeled documents as training set.

There are several issues in classification of documents. It is mainly big data problem ,High dimensionality that is large number of attributes which decreases the classifier performance. Another important feature is feature selection, to represent the features of document which can be done by different method which is binary representation and term frequency of occurrences.

In this paper different ways to classify news articles by using machine learning algorithms . We carry out a comparison of classification algorithms and evaluate a number of different feature sets with the goal of optimizing accuracy for the classification of news articles.

RELATED WORK

The model is related to Vandana Korde and C Namrata Mahender[1] on text classification and classifiers.The main goal of text mining is to enable users to extract information from textual resources and deals with the operations like, retrieval, classification (supervised, unsupervised and semi supervised) and summarization.

Natural Language Processing (NLP), Data Mining, and Machine Learning techniques work together to automatically classify and discover patterns from the different types of the documents . They compare different text classifier for their efficiency.

One more related research paper to my research was of Y. H. LI and A. K. Jain [2] says that paper investigate four different methods for document classification: the naive Bayes classifier, the nearest neighbour classifier, decision trees and a subspace method. These were applied to seven-class Yahoo news groups (business, entertainment, health, international, politics, sports and technology) individually and in combination.

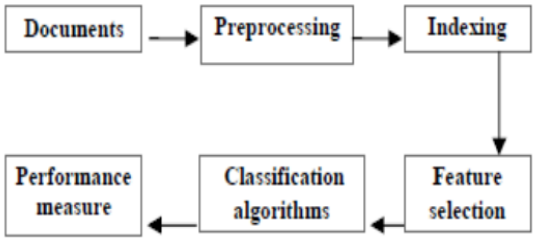
Mita K. Dalal and Mukesh A. Zaveri research paper [3] explains Automatic Text Classification is a semisupervised machine learning task that automatically assigns a given document to a set of pre-defined categories.

Mowafy M, Rezk A and El-bakry HM[4] explains An Efficient Classification Model for Unstructured Text Document by using multinomial naïve Bayes MNB with TF IDF and KNN and both for news articles.

The Proposed Model:

The proposed model presents comparsion of various machine learning methods for classification of telugu news articles and It show that support Vector Machine works well with an accuracy of 94 %. TF-IDF ,N grams and bag of words model for text extraction for more accurate text classification.

Document Collection: The first step of classification process includes collecting different types (format) of documents like .html, . pdf, .doc,etc. In this step, documents are collected, cleaned, and properly organized, the terms (features) are identified, and a vector space representation created.



Text Classification Model

Pre-processing:

The text document is represented in a word format in the preprocessing step. The preprocessing step consists of three steps mainly

Tokenization :

The contents of the file is tokenized into individual words.

Stop word removal :

Common words like articles, preposition's and pro-nouns etc are called stop words and they are removed.

Stemming word :

In the stemming method, stem/root of a word are identified. For example: the word connect, connected, connecting all together stemmed as "connect".

Feature extraction and selection:

It is the process of selecting a subset of the terms occurring in the training set and using only this subset as features in text classification. Feature extraction serves two main purposes. First, it makes training and applying a classifier more efficient by decreasing the size of the effective vocabulary. Second, feature selection often increases classification accuracy by eliminating noise features. A noise feature is one that, when included in the document representation, increases the classification error on new data.

For various types of text classification problems the most common baseline approach is the bag-of-words model.The bag-of-words (BOW) model considers each message as a set of words that each occurs a certain number of times. The representation of the document is entirely orderless, as each word is treated independently of the previous and upcoming word. As an example, a data set consisting of only two messages:

The weather is better than yesterday

The cat is better than the dog

	The	cat	is	better	than	dog	weather	yesterday
Vector one	2	1	1	1	1	1	0	0
Vector two	1	0	1	1	1	0	1	1

Bags of Words feature

N-gram model:

The N-gram model accounts for word order by counting sequences of words, where N is the number of words to include in a sequence. By including sequences of words, we are able to account for a deeper meaning in the sentences and capture more nuances in text. If using the first sample of the previous example and set N = 2,

Word N gram feature

the cat	cat is	is better	better than	than the	the dog
1	1	1	1	1	1

Character n-gram Model:

Character n-grams are similar to word n-grams, but instead of creating vector representations of words and word combinations, we create a vector representation based on characters and character combinations. As an example in the text one word N gram is one dog, one cat and 2-gram we would get the following feature vector (representing space-characters with on ne e_ _d do og g, _ o _c ca at

2	2	2	1	1	1	1	1	1

Classification and its Process:

Text classification is a fundamental task in document processing, whose goal is to classify a set of documents into a fixed number of predefined categories. Text categorization is the task of assigning a Boolean value to each pair {dj , ci} ϵ D × C, where D is a domain of documents and C = {c1,c2,.., ci} is a set of predefined categories. A value of T assigned to {dj , ci} indicates a decision to file dj under ci , while a value of F indicates a decision not to file dj under ci.

Logistic Regression:

In logistic regression on the bases of independent variables, discrete values are estimated (like 0/1 , yes/no ,true/false).By logit function the probability of occurrence of an event is predicted and the output values are between 0 and 1.

Testing and evaluating the model. In this step, the model is applied to the documents from the test set and their actual class labels are compared to the labels predicted. At this step the document labels are used for evaluation only, which is not the case at the validation step, where the class labels are actually used by the learning algorithm.

Naïve Bayes Classification Method:

Naïve Bayes is an technique that is used for the assignment of labels to problem instances in constructing classifiers methods, where feature values are represented by vectors, the class labels are taken from finite set. In this process, all the algorithms that are taken are considered with a common principle. This Naïve Bayes classifier assumes that the value of particular features is independent of other feature.

$$P(c_i \mid D) = \frac{P(c_i)P(D \mid c_i)}{P(D)}$$

$$P(D \mid c_i) = \prod_{j=1}^{n} P(d_j \mid c_i)$$

Where P(Ci)= $P(C = c_i) = \frac{Ni}{N}$

and P(dj|ci) = $P(d_j \mid c_i) = \frac{1 + N_{ji}}{M + \sum_{k=1}^{M} N_{ki}}$

Support Vector Machine:

The SVM is a method for training linear classifiers. It is based on statistical learning algorithms, it maps the documents into the feature space and attempts to find a hyperplane that separates the classes with largest margins. The SVM can be interpreted as an extension of the perceptron. It simultaneously minimizes the empirical classification error and maximizes the geometric margin. The working principle of SVM is to find out a hyper plane (linear/non-linear) which maximizes the margin. Maximizing the margin is equivalent to

$$\underset{w,b,\zeta_i}{\text{minimize}} \quad \frac{1}{2} w^T w + C\left(\sum_{i=1}^{N} \zeta_i\right)$$

$$\text{subject to} \quad y_i(w^T x_i - b) + \zeta_i - 1 \geq 0, \quad 1 \leq i \leq N$$

$$\zeta_i \geq 0, \quad 1 \leq i \leq N$$

Gradient Boost Tree:

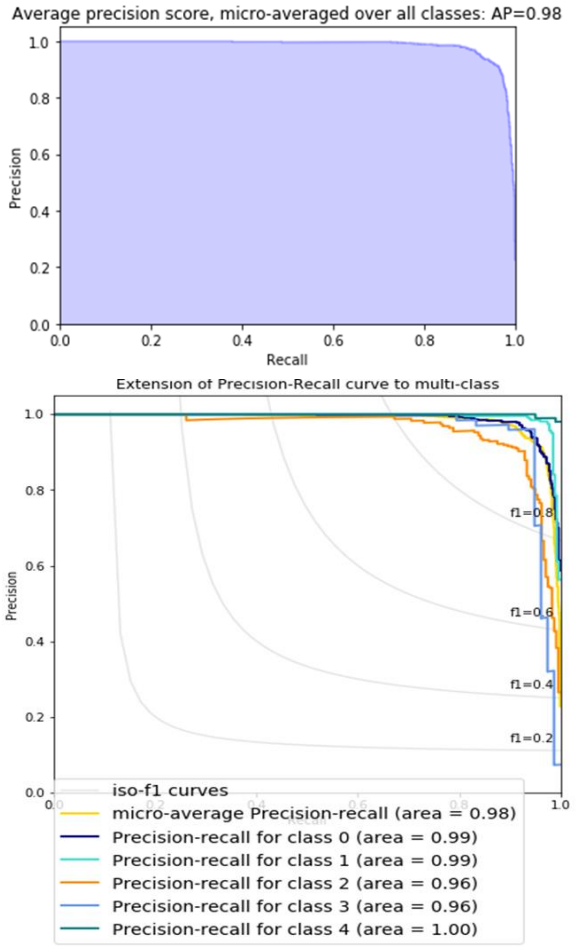
In this method binary tress are built i.e, the data is partitioned into two samples at each node. Now, if we limit the nodes complexity to 3, then root node along with two child node are called single split. After the determination of data that is partitioning, means for each partition are computed for deviations of observed values. So that, the next 3-node tree, that is fitted with these residuals, and further it is again partitioned, so that residual variance is reduced.

Using the two sets of training documents, we compared the four classification algorithms Naive Bayes classifier, Logistic Regression, Gradient Boosted Trees, Support Vector Machine models for text classification on test data sets. Here we keep human interpretation of the result (a human being classifying the documents into different categories after having a good knowledge of what exactly the dataset is as a gold standard so that we can compare the results of classifying algorithm with it and see how it behaves.

Experimental Results:

The following results are tested on telugu news articles. By using character N gram for SVM, Logistic regression, Gradient boost tree and Naïve Bayes accuracy and F1_score are measured. Out of all classifiers SVM provides better accuracy of 94.1

	SVM		Logistic regression		Gradient boost tree		Naïve Bayes	
	Accuracy	F1_score	Accuracy	F1_score	Accuracy	F1_score	Accuracy	F1_score
Character N gram 1-2	91.74	0.94	85.9	0.85	82.4	0 .83	80.09	0.80
Character N gram 1-3	93.2	0.94	86.9	0.90	84.2	0.84	83.17	0.83
Character N gram 1-4	94.1	0.94	87.3	0.91	85.1	0.85	84.2	0.83



CONCLUSIONS

In 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. Measures have been used, like accuracy and F1_Score are calculated.

REFERENCES

[1] Mita K. Dalal, Mukesh A. Zaveri "Automatic Text Classification: A Technical Review", International Journal of Computer Applications (0975 – 8887),Volume 28– No.2, August(2011).

[2] K. Naleeni, Dr.L.Jaba Sheela,"Survey on Text Classification", International Journal of Innovative Research in Advanced Engineering (IJIRAE), Volume 1 Issue 6 July (2014).

[3] Kratarth Goel, Raunaq Vohra, Ainesh Bakshi, "A Novel Feature Selection and Extraction Technique for Classification", IEEE International Conference on Systems, Man, and Cybernetics, October 5-8,(2014).

[4] Mowafy M*, Rezk A and El-bakry HM "An Efficient Classification Model for Unstructured Text Document by using multinomial naïve Bayes MNB with TF IDF and KNN and both for news articles"American Journal of Computer Science and Information Technology(2018).

[5] Sang-Bum Kim, Kyoung-Soo Han, Hae-Chang Rim, and Sung Hyon Myaeng, "Some Effective Techniques for Naïve Bayes Text Classification", IEEE transactions on Knowledge and Data Engineering, VOL. 18, NO. 11, November (2006).

[6] Ioan Pop,"An Approach of the Naïve Bayes Classifier for the document classification", General Mathematics Vol. 14, No. 4 (2006).

[7] George Tsatsaronis ,Vicky Panagiotopoulou, "A Generalized Vector Space Model for Text Retrieval based on Semantic Relatedness", Association for Computational Linguistics, Athens, Greece, 2 April (2009).

[8] Y.H. Chen,Y.F. Zheng, J.F. Pan, N. Yang,"A hybrid text classification method based on K-congenernearest- neighbors and hypersphere support vector machine", International Conference on Information Technology and Applications, (2013).

[9] Tam, V., Santoso, A., & Setiono, R. , "A comparative study of centroid-based, neighborhood-based and statistical approaches for effective document categorization", Proceedings of the 16th International Conference on Pattern Recognition, pp.235–238, 2002.

[10] S.L.Ting,W.H.Ip, Albert.H.C.Tsang ,"Is Naïve Bayes a Good Classifier for Classification?",International Journal of Software Engineering and Its Applications, Vol. 5, No. 3, July, (2011)