Text Categorization with Ensemble Methods

Bhavya Gupta

[bhavya8299460915@gmail.com](mailto:bhavya8299460915@gmail.com)

Department of Computer Science and Engineering,

University Institute of Engineering and Technology,

CSJM University, Kanpur

Dr. Alok Kumar

akumar.uiet@gmail.com

Department of Computer Science and Engineering,

University Institute of Engineering and Technology,

CSJM University, Kanpur

**ABSTRACT:** - The objective is to compare the performance of four different text classifiers and use ensemble methods to produce optimal predictive model. In ensemble method (classifier combination method) the predictions of each base classifier are combined into one predictive model to decrease the variance, to decrease the bias and to improve the predictions. We have implemented four different text classifiers using Naïve Bayes, Logistic Regression, Random forest and neural network algorithms. We have experimented with different encoding schemes (Normal count, log count, TFIDFvectorizer etc). In our research we find that the performance of ensemble model is better than the performances of multiple models that we had made for the text categorization. The average accuracy of Ensemble model is 98.45%.In our experiment, we have collected data from kaggle.com.

**Keywords:** Ensemble Methods, Naïve Bayes, Logistic Regression, Random Forest, Neural Networks.

**INTRODUCTION**

With the advent of the Internet, the rate of production of data is increasing exponentially. In today’s scenario the data generated by humans is very huge. Over 2.5 quintillion bytes of data are created every single day. This data would be a waste if it cannot be used to acquire some knowledge from it. So, there are many techniques that are applied to this huge data to learn some trends, behavior and acquire some kind of information from it. Text analysis is one of the ways of data processing. As we are gathering a large amount of data in the form of text such as web clicked data, Twitter data, Facebook data, and the data gathered from various other sites and surveys. So, this data can be used to gather new trends if some processing is performed in it by using some machine learning algorithms. In this paper, we are performing text categorization. Text categorization is one of the ways of data processing. Text Categorization is a supervised machine learning problem in which a dataset is provided which consist of some statements and corresponding to each statement a label is given. In this paper, a very high accuracy is achieved by using the power of various classification algorithms. In our research, we have used Naive Bayes, Logistic Regression, Random Forest and Neural Network for classification and then used their prediction for prediction of final output. For preprocessing of data, we have used TFIDF Vectorizer and Countvectorizer.

# MOTIVATION

With the advent of the Internet a lot things has become simpler. But it has its negative aspect also. It is used by some anarchists, religious traitors and politicians to inflame instability in the society. So, Text Categorization can help a lot in solving this problem. By classifying this previously and taking action against them. Along With text categorization can also help in speeding web searches if it learns to classify web pages and shows only those web pages that belong to the query. Therefore, this topic has a very wide implementation in the real world. So, it requires algorithms that can provide good accuracy. In this paper, the accuracy of the existing algorithms is boosted by using ensemble method.

# Related Work

The accessible reviews mostly describe and focus on the following five elements of the text classification process: (1) document pre-processing, i.e. tokenization, stop-word removal, and stemming or lemmatization, (2) document modeling, i.e. representing a document in an appropriate form, to be processed by a machine learning algorithm, (3) feature selection and projection, (4) machine learning algorithm utilization to construct a classification model or function, and (5) quality indicators and evaluation methods.

Aas and Eikvil have previously enumerated and de- scribed the text classification steps, namely, pre-processing, vector space model creation, dimensionality reduction (feature selection and projection), training of a classification function, and performance measurement. In their work, text classification was used to present several schemas of feature weighting, e.g. Boolean, term frequency, inverse document frequency, and entropy. Moreover, the authors explained three feature selection methods, namely, document frequency thresholding, information gain, and χ2 -Statistic, and one feature projection method, namely, latent semantic indexing (LSI). In addition, they summarized and elucidated six machine learning methods: Rocchio’s algorithm, naive Bayes, k- nearest neighbor, decision tree, support vector machine (SVM), and ensemble learning, including bagging and boosting algorithms. Furthermore, they described the performance measures for binary, multi-class, and multi-label classification tasks. They also provided a Reuters-21578 dataset, suitable for different classification experiments.

Aggarwal and Zhai described text processing similarly to Aas and Eikvil (1999), but they provided more examples. Additionally, the authors presented a more in depth discussion on each element of the text classification process. Their paper starts with a description of the classification problems, software, and examples of the domains in which text classification is commonly used. In the article, the authors introduce (1) feature selection methods, including the Gini index, information gain, mutual information, and χ2 -statistic, and (2) feature projection methods, such as different types of LSI, supervised clustering for dimensionality reduction, linear discriminant analysis, and generalized singular value decomposition. Second, they describe the following different types of classification learning algorithms: decision tree, rule based classifiers, naive Bayes (multivariate Bernoulli and multinomial models), SVM, regression-based classifiers, neural network, and proximity-based classifiers (k-nearest neighbors, Rocchio’s). In addition, they discuss ensemble learning techniques, including simple committees, boosting, and bagging. Finally, Aggarwal and Zhai (2012) explain the measures of accuracy of the classification process. We also notice some implicit and explicit observations made in their paper regarding the classification techniques. These observations are related to, for example, different types of classification tasks and their solutions and performance of the linear classifiers. Moreover, it is worth mentioning that Aggarwal and Zhai (2012) address the interesting problems of linked and web data classification

# SYSTEM DESIN

There are two main components of our system i.e. Data Preprocessing and Ensemble learning. Data Processing is one of the main steps of any machine learning algorithms to remove unwanted text. In Ensemble learning, multiple learning algorithms are used to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone. Flow chart of proposed system is shown in figure 1.



Figure 1: Flow chart of proposed system

1. **Data Preprocessing**

In data preprocessing the noise present in our dataset is removed. By removing the noises we can build a good model that will provide a good accuracy. In this paper, the dataset used consist of statements and the labels associated with them. Cleaning of statements is performed by removing all the stopwards that will not contribute to our result. Then the stemming process is done to further reduce the vocabulary size. Stemming is the process in which we remove similar appearance word with root word. Dimension Reduction is the process of transforming a set of data having large number of dimensions into data with less number of dimensions confirming that it will provide same information briefly. These methods are mainly used while solving machine learning problems to obtain fine details for a classification or regression function. Following are the advantages of Dimension Reduction

* + - Data compression and reduction of the storage space.
    - It reduces the time required for performing similar computation. Less dimensions make the algorithm work efficiently which was not running properly for large dimension set.
    - Improves the model performance by taking care of multi co linearity. Reductant features such as value with two different unit is removed.
    - Data visualization becomes easier by reducing the dimensions of data to 2D or 3D.

There are two ways of Dimension Reduction:

1. Feature Selection: Process of removing irrelevant or reductant features from the dataset.
2. Feature Extraction: Process of generating a new, smaller set of features that present most of the useful information.

In this research TFIDF Vectorizer and Count Vectorizer is used to reduce the features.

*TF-IDF Vectorizer :-*The grail of using tf-idf is to scale down the impact of tokens that occur very frequently in a given corpus and that are hence empirically less informative than features that occur in a small fraction of the training corpus. TF-IDF score of a term is computed as.

tf-idf(d, t) = tf(t) \* idf(d, t)

*Countvectorizer :-*The countvectorizer come up with a simple way to both tokenize a collection of text documents and set up a vocabulary of known words, but also to encode new documents using that vocabulary.

We can make use of it as follows:-

1. Initiate the instance of Countvectorizer class.
2. In order to acquire the knowledge of a vocabulary from one or more than one document,call the fit() function.
3. On one or more documents, we call the transform() function as entailed to encode each as a vector.

After applying the above processes we have the encoded documents in the vector format.

1. **Ensemble learning**

For improving the accuracy of model, we are gathering results from all the models and deciding the output based on category that comes more times. In this research, we have used Naïve Bayes, Logistic Regression, Random Forest and Neural Network classifiers to finalize class of a text.

*Bayes Classification:*A Naive Bayes classifier is a probabilistic machine learning model that’s used for classification task. The crux of the classifier is based on the Bayes theorem.

P(A/B) = (P(B/A)P(A))/P(B)

Using Bayes theorem, we can find the probability of A happening, given that B has occurred. Here, B is the evidence and A is the hypothesis. The assumption made here is that the predictors/features are independent. That is presence of one particular feature does not affect the other. Hence it is called naive.

*Logistic Regression Classification:* Logistic Regression is a Machine Learning algorithm which is used for classification problems; it is a predictive analysis algorithm and based on the concept of probability. We can call a Logistic Regression a Linear Regression model but the Logistic Regression uses a more complex cost function, this cost function can be defined as the ‘Sigmoid function’ or also known as the ‘logistic function’ instead of a linear function. The hypothesis of logistic regression tends it to limit the cost function between 0 and 1. Therefore linear functions fail to represent it as it can have a value greater than 1 or less than 0 which is not possible as per the hypothesis of logistic regression.

*Random Forest Classification:* A random forest is a data construct applied to machine learning that develops large numbers of random decision trees analyzing sets of variables. This type of algorithm helps to enhance the ways that technologies analyze complex data and the procedure is to be visible in figure 2. Random forest classifier creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object.

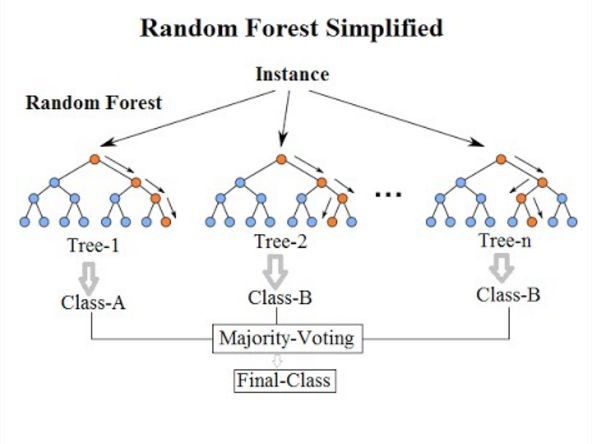


Figure 2: Random Forest

# *Neural Network:* Artificial neural networks (ANN) or connectionist systems are computing systems that are inspired by, but not identical to, [biological neural networks](https://en.wikipedia.org/wiki/Biological_neural_network) that constitute animal [brains](https://en.wikipedia.org/wiki/Brain). Such systems "learn" to perform tasks by considering examples, generally without being programmed with task-specific rules. For example, in [image recognition](https://en.wikipedia.org/wiki/Image_recognition), they might learn to identify images that contain cats by analyzing example images that have been manually [labeled](https://en.wikipedia.org/wiki/Labeled_data) as "cat" or "no cat" and using the results to identify cats in other images. They do this without any prior knowledge of cats, for example, that they have fur, tails, whiskers and cat-like faces. Instead, they automatically generate identifying characteristics from the examples that they process. A multilayer perceptron to be trained for classification that is depicted in the figure 3.

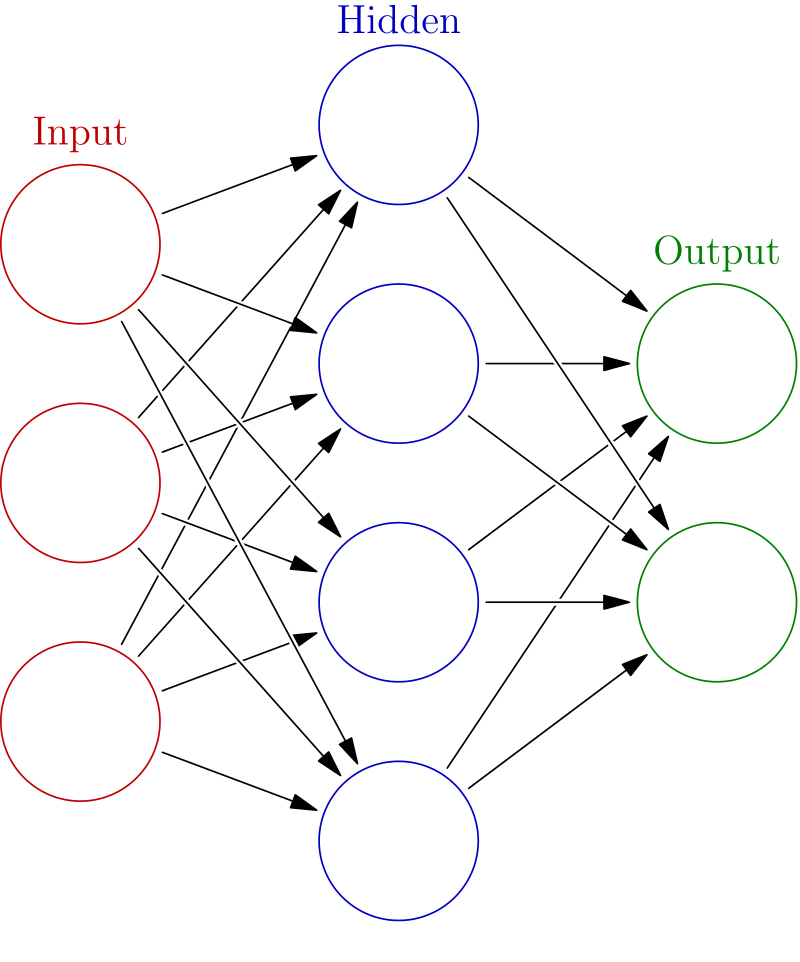


Figure 3: Multilayer perceptron

# RESULT AND CONCLUSION

In this paper, the various approaches of text categorization are explored. It has been observed that the supervised learning methods mainly Naïve Bayes (NB), Logistic Regression, Random Forest and Neural Networks are suitable for text classification. After getting the result from all the algorithms, we make use of ensemble method that is also a method of machine learning which then combine the result of all these algorithms and give the improved output. Performance of all used algorithm is compared in table 1. On the basis of this research, we can conclude that performance of a machine learning system can be enhanced by Ensemble learning.

|  |  |  |  |
| --- | --- | --- | --- |
| **Sr. No.** | **Algorithms used** | **Accuracy** | **Statements** |
| **1)** | Naive Bayes Classification | 92.00% | In this we make use of bayes theorem to predict the model. |
| **2)** | Logistic Regression | 95.00% | Similar to naive Bayes, logistic regression is a probabilistic classifier that makes use of  logistic function. |
| **3)** | Random Forest | 90.00% | Random forests is a set of multiple decision trees. |
| **4)** | Neural Networks | 89.00% | It classify phrases into a set of predefined categories using ANN. |
| **5)** | Ensemble method | 97.00% | In this we gather the results from all the algorithms to get the most accurate result. |

# REFERENCES

1. B. Mahalakshmi, Dr. K. Duraiswamy, “An Overview of Categorization Techniques” 2249-6645, International Journal of Modern Engineering Research (IJMER). Oct 2012.
2. Bhumika, Prof. Sukhjit Singh Sehra, Prof. Anand Nayyar, “A Review Paper on Algorithms Used for Text Categorization”, 2319- 4847, International Journal of Application or Innovation in Engineering Technology & Management, March 2013.
3. Bijal Dalwadi, Vishal Polara, Chintan Mahant, “A Review: Text Categorization for Indian Language”, 2349-4476, International Journal of Engineering Technology, Management and Applied Sciences, March 2015.
4. István Pilászy, “Text Categorization and Support Vector Machines”
5. Meenakshi, Swati Singh, “Review Paper on Text Categorization Techniques” ISSN: 2348-8387, SSRG International Journal of Computer Science and Engineering(SSRG-IJCSE)-EFES April 2015.
6. Munirul Mansur, NaushadUzZaman , Mumit Khan, “Analysis of NGram Based Text Categorization for Bangla in Newspaper Corpus”.
7. Nadimpalli V Ganapathi Raju et. al., “Automatic Information Collection & Text Classification for Telugu Corpus using K-NN”, 2231-1009, International Journal of Research in Computer Application & Management, November-2011.
8. Nidhi, Vishal Gupta, “Punjabi Text Classification using Naïve Bayes, Centroid and Hybrid Approach”, DOI: 10.5121/csit.2012.2421.
9. S. Niharika, V. Sneha Latha, D. R. Lavanya, “A Survey on Text Categorization”, 2231-2803, International Journal of Computer Trends and Technology, 2012.