[NC-Rase 18]
DOI: 10.5281/zenodo.1483957

ISSN 2348 – 8034
Impact Factor- 5.070

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

# CLUSTERING AND CLASSIFICATION TECHNIQUES USING TEXT MINING

**VarshaC. Pande[*1], Dr. Harshala B. Pethe[2] & Dr. Abha. S. Khandelwal[3]**

[*1]Research Scholar, Department of Electronics and Computer Science, RTMNU, Nagpur, (MH), India,
[2]Assistant Professor, Department of Computer Science, Dr. Ambedkar College, Dikshabhoomi, Nagpur, (MH), India [3]Former-HOD, Department of Computer Science, Hislop College, Member BOS (CS), RTM Nagpur University, Nagpur,( MH), India

## ABSTRACT

The text is nothing but the combination of characters. Therefore, analyzing and extracting information patterns from such data sets are more complex. Several methods have been proposed for analyzing such texts and extracting information. Data mining, a specific area named text mining is used to classify the huge semi structured or unstructured data needs proper clustering. Maximum text documents involves fast retrieval of information, arrangement of documents, exploring of information from the documents. Declaration of text input data and classification of the documents is a complex process.

Text Clustering is an unsupervised method in which no input out patterns is predefined. This method is based upon the idea of dividing the similar text into the same cluster. Individual cluster consists of number of records. The clustering is thought better if the contents of documents of intra cluster are more alike than the contents of inter-cluster documents.

Classificationis used to find out in which group each data instance is related within a given dataset. It is used for classifying data into different classes according to some constrains. Several major kinds of classification algorithms including C4.5, ID3, k-nearest neighbor classifier, Naive Bayes, SVM, and ANN are used for classification. This paper describes the comparative study of clustering and Classification Algorithms.

***Keywords:*** *Data mining, Text mining, Classification, Clustering and Rapid Miner.*

## I. INTRODUCTION

Text Mining [1] is the process of extracting useful information or patterns from the unorganized (unstructured) text that are from various sources. As the text is in unorganized form, it is quite difficult to handle it. For finding interesting information from the natural language text is the main purpose of text mining. The text mining process is shown in below figure 1:

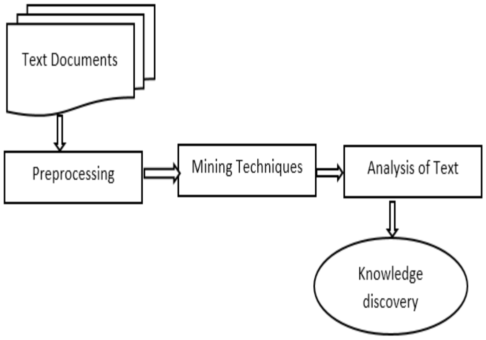


***Figure1: Framework of Text Mining***

Here we present the approaches for the analysis of tasks preprocessing, classification and clustering.

Step I- Pre-processing Text: As compare to natural languages documents, Mining from a pre-processed document is easy. Thus, pre-processing of documents is an important task during text mining process before applying any text mining technique. As Text documents can be represented as - a bag of words on which different text mining methods. To reduce the dimensionally of the texts words, appropriate methods such as filtering and stemming are used. Filtering techniques remove those words from the set of all words that do not give relevant information; stop word filtering is a conventional filtering method. After this step is applied, every word is represented by its root word.

Step II- Mining Technique: In this step the selected algorithm is applied to text in order to process the document. Here, the clustering and Classification and classification algorithms are used.

Step III - Analysis of Text: For knowledge discovery purpose, outputs which are coming from initial stage are analyzed here. For this purpose, various tools such as link discovery tool can be used. Here the unstructured text has been converted into some meaningful information from which one can make decisions.

## II. CLASSIFICATION ALGORITHMS

Classification [2, 3] means turning over a document or object to one or more classes. This may be done manually or algorithmically. Classification is done mainly based on attributes, behavior or subjects.

Classification [4] is a mining technique that assigns categories to a collection of data in order to aide in more accurate predictions and analysis [5], is known Decision Tree, classification is a method intended to make the analysis of very large data sets effective. To create an effective set of classification rules which answers a query, makes decision based on the query and predicts the behavior. To begin with a set of training data sets are created with certain set of attributes or outcomes.

The Classification [6] problem can be specified as a training data set consisting of records. Each record is identified by unique record id, and consist of fields corresponding to the attributes. The continuous attribute is an attribute with a continuous domain and an attribute with a finite domain of discrete values is called a categorical attribute. One of the categorical attribute is the classifying attribute or class and the value in its domain are called class labels.The main objective of the classification algorithm is, how to set of attributes reaches its conclusion.

### *K-Nearest Neighbor (KNN)*

K-Nearest Neighbor [4]classifier is an Algorithm which is based on a distance function for pairs of observations, such as the Euclidean distance or Cosine. In this hypothesis, k nearest neighbors of a training data is computed first. Then the similarities of one sample from testing data to the k nearest neighbors are aggregated according to the class of the neighbors, and the testing sample is assigned to the most similar class. KNN is nonparametric lazy learning algorithm. A technique is nonparametric, it means that it does not make any assumptions on the underlying data distribution.

### *Naive Bayes Classifier*

***Naive Bayes***[7]is used to deal with the problem of document classification by a deceptively simplistic model. The Naive Bayes approach is applied in Flat (linear) and hierarchical manner for improving the efficiency of classification model. It has been found that Hierarchical Classification technique is more effective than Flat classification. It also performs better in case of multi-label document classification. Bayesian classifier is a statistical classifier as well as a supervised learning method. It will predict class membership probabilities. It calculates explicit probabilities for hypothesis and it is robust to noise in input data. When Bayesian classifier is applied to large datasets, it shows high accuracy and speed. Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes.

### *Decision tree*

Decision tree [8]is classification algorithm in which there are several popular decision algorithms such as Quinlan's ID3, C4.5, C5, and CART [9]. A decision tree is a flow-charting like structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label [10]. This technique separates observation into branches to construct tree on repetition basis. In most cases, tree classifiers perform classification in two stages: tree-growing and tree-pruning. The tree-growing is top down approach. In this stage, the tree is split in a recursive manner called recursive partitioning. It is completed when the subset at a node has all the same value of the target variable, or when splitting no longer adds value to the predictions. In the tree-pruning, the tree will be fully grown, fully grown tree is cut back to avert over fitting data and this way it improves the correctness of the tree in bottom up manner. This technique is used to improve the estimate and correctness of the algorithm by minimizing the over fitting. Decision tree is widely used in various areas because it is strong enough for data distribution.

***Random Forest***

Random forests [4]is an ensemble learning method. It is one of the accurate learning algorithm. The basic concept of the algorithm is to build many small decision-tree and then merging them to form a forest. It is computationally easy and cheap process to build many such small and weak decision trees. So such decision trees can be formed in parallel and then it can be combined to form a single and strong forest. The algorithm for random forests uses the common technique of bootstrap bagging. Given a training set S = {(x1, y1),……,(xn, yn)}, bagging repeatedly (B times) selects a random sample from the training set and construct trees to fit these samples. This procedure leads to better performance that while the predictions of a single tree are highly sensitive to noise in its training set, the average of many trees is not, as long as the trees are not correlated.

***SVM Based Classification***

Support vector machines [11]are based on the Structural Risk Minimization principle [12] from computational learning theory. The idea of structural risk minimization is to find a hypothesis h. The true error of h is the probability that h will make an error on an unseen and randomly selected test example. An upper bound can be used to connect the true error of a hypothesis h with the error of h on the training set and the complexity of H (measured by VC-Dimension), the hypothesis space containing h [12]. Support vector machines find the hypothesis h which minimizes this bound on the true error by effectively and efficiently controlling the VC-Dimension of H. SVMs are very universal learners, in their basic form, SVMs learn linear threshold function. Nevertheless, by a simple \plug-in" of an appropriate kernel function, they can be used to learn polynomial classifiers, radial basic function (RBF) networks, and three-layer sigmoid neural nets. One remarkable property of SVMs is that their ability to learn can be independent of the dimensionality of the feature space. SVMs measure the complexity of hypotheses based on the margin with which they separate the data, not the number of features. This means that we can generalize even in the presence of very many features, if our data is separable with a wide margin using functions from the hypothesis space.

The same margin argument also suggest a heuristic for selecting good parameter settings for the learner.  The best parameter setting is the one which produces the hypothesis with the lowest VC-Dimension. This allows fully automatic parameter tuning without expensive cross-validation.

***Text Clustering***

Clustering [13]is one of the commonly used unsupervised learning methods for analyzing the context of text data in natural language form [14].Clustering [15]is the process of grouping or classifying objects based on information obtained from the data describing the relationship among objects in principle to maximize the similarities among members of the same class and to minimize the similarities among the class or cluster [16]. It is a mathematical approach in collecting and segmenting similar documents into clusters. It helps trim down the volume of unstructured text and provide a simpler understanding and thematic structure of the data. It also provides the keywords in each cluster that is useful in extracting valuable insights, hence, customer sentiments can be summarized using these keywords.

Clustering is concerned with grouping objects together that are similar to each other and dissimilar to the objects belonging to other clusters.For our analysis have chosen Random Clustering algorithm to cluster documents.

This algorithm performs a random flat clustering of the given Dataset. Please note that this algorithm does not guarantee that all clusters will be non-empty. It creates a cluster attribute in the resultant Dataset. It is important to note that this algorithm randomly assigns examples to clusters.

## III. RESULT & DISCUSSION

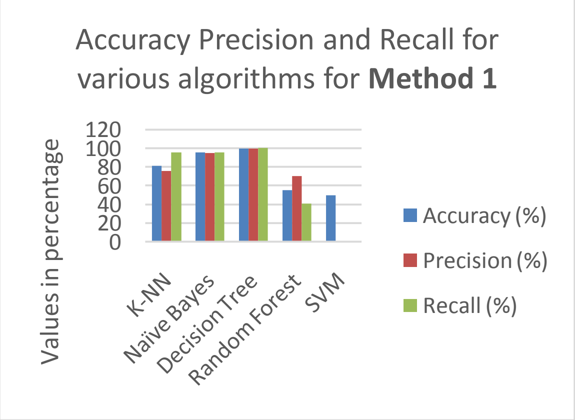
For the analytical study we have taken the two folders i.e. **Sports Baseball** and **Sports Hockey** from **Mini 20 News Group Dataset.** The values for **Accuracy, Precision**,**Recall and Execution Time** for Random Clustering and Classification Algorithms i.e. K-NN, Naïve Bayes, Decision Tree, Random Forest and Support Vector Machine (SVM) are calculated. Two methods are considered in this paper. In **Method 1** we have executed Classification algorithm first then the clustering Algorithm and in **Method 2** we have executed Clustering algorithm first then the Classification Algorithm.

The following Table1 shows the values for metrics, when method 1 is executed.
**Tables**:

*Table 1. Performance of Method 1*

| Metrics<br>Algorithm | Accuracy (%) | Precision (%) | Recall (%) | Time (min & sec) |
|---|---|---|---|---|
| **K-NN** | **80.35** | **74.77** | **94.60** | **5:01** |
| **Naïve Bayes** | **94.85** | **94.82** | **94.90** | **1:26** |
| **Decision Tree** | **99.70** | **99.41** | **100** | **1:47** |
| **Random Forest** | **55.15** | **78.77** | **38.10** | **4:17** |
| **SVM** | **50.00** | **Unknown** | **0.00** | **2:44** |

From Graph 1 it is clear that, Decision Tree algorithm is better among all the five algorithms.

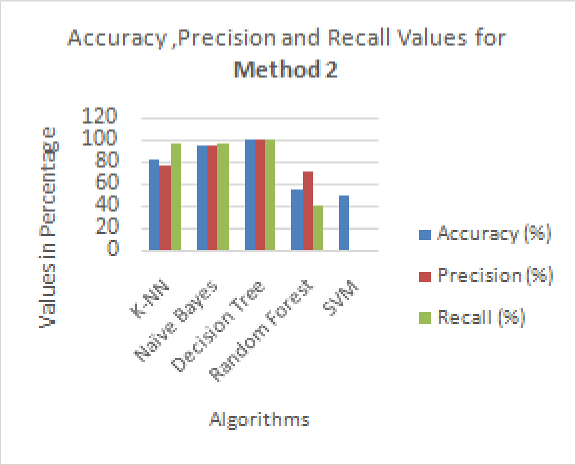


*Graph 1: Results for Method 1*

The following Table 2 shows the values of metrics, when we execute Method 2.

*Table 2. Performance of Method 2*

| Metrics Algorithm | Accuracy (%) | Precision (%) | Recall (%) | Time (Min & Sec) |
|---|---|---|---|---|
| K-NN | 81.40 | 75.58 | 95.60 | 5:8 |
| Naïve Bayes | 95.30 | 95.00 | 95.70 | 1.34 |
| Decision Tree | 99.70 | 99.41 | 100 | 1:43 |
| Random Forest | 55.10 | 70.2 | 40.84 | 4:31 |
| SVM | 50 | Unknown | 0.0 | 2:44 |



***Graph 2: Results for Method 2***

From Graph 2 it is clear that, Decision Tree algorithm is better among all the five algorithms.

## IV. CONCLUSION

This paper deals with various Classification and Clustering techniques used in TextData mining. Data Mining is a wide area that integrates techniques from variousfields including machine learning, artificial intelligence,statistics and pattern recognition, for the analysis of largevolumes of data. Classification methods are typically strongin modeling interactions, these classification algorithms are implemented on 20 News Group dataset. In this paper, we have compared and analyzed the classification and clustering algorithms. The clustering and classification algorithms are executed in sequence and analyzed. We observed the results in Method 2 is better than Method 1. Classification algorithms such as K-NN, Naïve Bays, Decision Tree, Random Forest and SVM compare with the results for Decision Tree algorithm is better in terms of both the Method – 1 & 2 Accuracy (99.70), Precision (99.41) and Recall (100).

## REFERENCES

1. *N. VenkataSailaja,L. Padmasree and N. Mangathayaru,"Survey of Text Mining Techniques, Challenges and their Applications"International Journal of Computer Applications (0975 – 8887), Volume 146 – No.11, July 2016.*
2. *A. Purohit, D. Atre, P. Jaswani and P.Asawara,"Text Classification in Data Mining", International Journal of Scientific and Research Publications, Volume 5, Issue 6, June 2015 ISSN 2250-3153.*
3. *A. Hossain, R.Mamunur and M. Chowdhury, "A New Genetic Algorithm Based Text Classifier," In Proceedings of International Conference on Computer and Information Technology, NSU, pp. 135-139, 2001.*
4. *S.Ponmani, R. Samuel, P.VidhuPriya,"Classification Algorithms in Data Mining – ASurvey", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)Volume 6, Issue 1, January 2017, ISSN: 2278 – 1323.*
5. *RaúlVicen-Bueno, Rubén Carrasco-Alvarez, MaríaPilarJarabo-Amores, José CarlosNieto-Borge, and Enrique Alexandre "Detection of Ships in Marine Environments bySquare Integration Mode and Multilayer Perceptrons" IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT, VOL. 60, NO. 3, pp 712-724,Mar 2011.*
6. *S. Ghosh, S. Roy, and S. Bandyopadhyay, "A tutorial review on Text Mining Algorithms", International Journal of Advanced Research in Computer and Communication EngineeringVol. 1, Issue 4, June 2012,ISSN : 2278 – 1021.*
7. *E. Jadon and R. Sharma, "Data Mining: Document Classification using Naive Bayes Classifier", International Journal of Computer Applications (0975 – 8887)Volume 167 – No.6, June 2017.*
8. *C. Shah and A.Jivani, "Comparison of Data Mining Classification Algorithm for Breast Cancer Prediction", Research Gate Conference Paper, July 2013DOI: 10.1109/ICCCNT.2013.6726477.*
9. *Delen, D., Analysis of cancer data: a data mining approach. Expert Systems, 26: 100–112. doi: 10.1111/j.1468-0394.2008.00480.x(2009).*
10. *J. Han and M.Kamber, "Data Mining Concepts and Techniques", third edition, Morgan Kaufmann Publishers an imprint of Elsevier.*
11. *T. Joachims,"Text Categorization with Support VectorMachines: Learning with Many RelevantFeatures".*
12. *V. Vapnik, "The Nature of Statistical Learning Theory", Springer, NewYork, 1995.*
13. *A. Halibas, A. Shaffi and M. Varusai Mohamed, "Application of Text Classification and Clusteringof Twitter Data for Business Analytics" Research Gate, Conference Paper · March 2018DOI: 10.1109/MINTC.2018.8363162.*
14. *N. Yussupova, M. Boyko, and D. Bogdanova, "A Decision SupportApproach based on Sentiment Analysis Combined with Data Mining forCustomer Satisfaction Research," Int. J. Adv. Intell. Syst., vol. 1&2,2015.*
15. *T. Winarti, J. Kerami and S.Arief, "Determining Term on Text Document Clustering using Algorithm of Enhanced Confix Stripping Stemming", International Journal of Computer Applications (0975 – 8887), Volume 157 – No 9, January 2017.*
16. *Arai, K., Barakbah, A. R 2007,"Hierarchical K-Means: an algorithm for centroids initialization for K-Means", the Faculty of Science and Engineering, Saga University, Vol. 36, No.*