Scientific Journal Impact Factor (SJIF): 1.711

International Journal of Modern Trends in Engineering and Research

IJMTER

ISSN (ONLINE) : 2349 - 9745

ISSN (PRINT) : 2393 - 8161

Efficient Machine Learning Classifiers for Automatic Information Classification

Dr. S. Vijayarani[1], Ms. N. Nithya[2]

[1]Assistant Professor, Department of Computer Science, Bharathiar University .

[2]M. Phil Research Scholar, Department of Computer Science, Bharathiar University.

Abstract - As the technology keeps on developing to tremendous heights, the maintenance of a huge data becomes harder and harder in day to day life. These huge data cannot be used until it is in an understandable manner. So the method to handle these huge data is done by using data mining methods. Data Mining is defined as extraction of unseen information from the huge set of data. The data mining tasks are classification, prediction, outlier analysis, clustering, association rules, correlation analysis and time series analysis. One of the important domains in data mining, which handles the text data, is called as Text Mining. Text mining generally refers to the process of extracting interesting information and knowledge from unstructured text data. Text mining seeks to extract useful information from unstructured textual data through the identification and exploration of interesting patterns. This research paper uses the concept of machine learning. It pre-processes the dataset, searches for keyword from the dictionary, trains the machine using the algorithms SVM and naïve bayes algorithms. The proposed method is tested and validated using the Annexure II journal list dataset. From the experimental results we analyze that the SVM algorithm produces better results than the Naïve Bayes. The result of the proposed method is used to improve the efficiency of the real time application in both the government and the private agencies.

Keywords - Text Mining, Pre-processing techniques, Classification, Machine learning, SVM and Naïve Bayes.

I. INTRODUCTION

In the field of Information technology, we have a huge amount of data available that need to be turned into useful information. Data mining is used to mine the knowledge from data. The data can be of various types like text, audio, video, graphics, web pages, etc. The various applications such as market analysis, fraud detection, customer retention, production control, science exploration, etc. The other applications of data mining are sports, astrology and Internet Web Surf-Aid.

Text mining is nothing but finding the attractive patterns in large textual datasets. Where the word attractive refers to non-trivial, previously unknown, unlabeled data and hidden information [2]. Some of the application of text mining is enhancing the web search, mining the bibliographic data, document classification, topic categorization, newsgroup categorization, sentiment classification, etc. The data analysis is done for extracting the models and predicting the future data trends with the help classification [1]. Classification predicts categorical and prediction models which can predict continuous valued functions. It helps us to provide a better understanding of large data in an easy manner. For example, we can build a classification model to categorize bank loan applications as either safe or risky based on the customer income. The text classification involves two steps, namely [1],

(I) **Building the classifier** - This step is the learning step or the learning stage. In this step the classification algorithms build the classifier. The classifier is built from the training set made up of database tuples and their associated class labels. Each tuples that constitutes the training set is referred to a class or category.

(II) **Using the classifiers for classification -** In this step the classifier is used for classification. Here the test data is used to estimate the correctness of classification rules. The classification rules can be applied to the new data tuples if the exactness is considered acceptable.

The problem statement of this research work is to categorize the text information in the table accordingly to their disciplines. The categorization technique is done by training the machine, once the machine is trained with some instances, and then they can learn and categorize the new arriving instances to the table. By training the machine we can overcome the drawbacks like time consumption for manually classifying the data, human stress and incorrect categorization. The machine learning algorithms used in this research work are SVM and Naïve Bayes algorithm. This research work is tested using the synthetic dataset which is extracted from Annexure II in the Anna university website, it is pre-processed using the techniques stemming and removal of stop words, searches for a keyword in the dictionary and then it classifies the text information using the machine learning algorithms namely SVM and Naïve Bayes.

The remaining portion of the paper is organized as follows. Section 2 gives the related works. Section 3 describes the proposed methodology. Experimental results and screenshots are discussed in Section 4. Conclusions are given in Section 5.

II. LITERATURE REVIEW

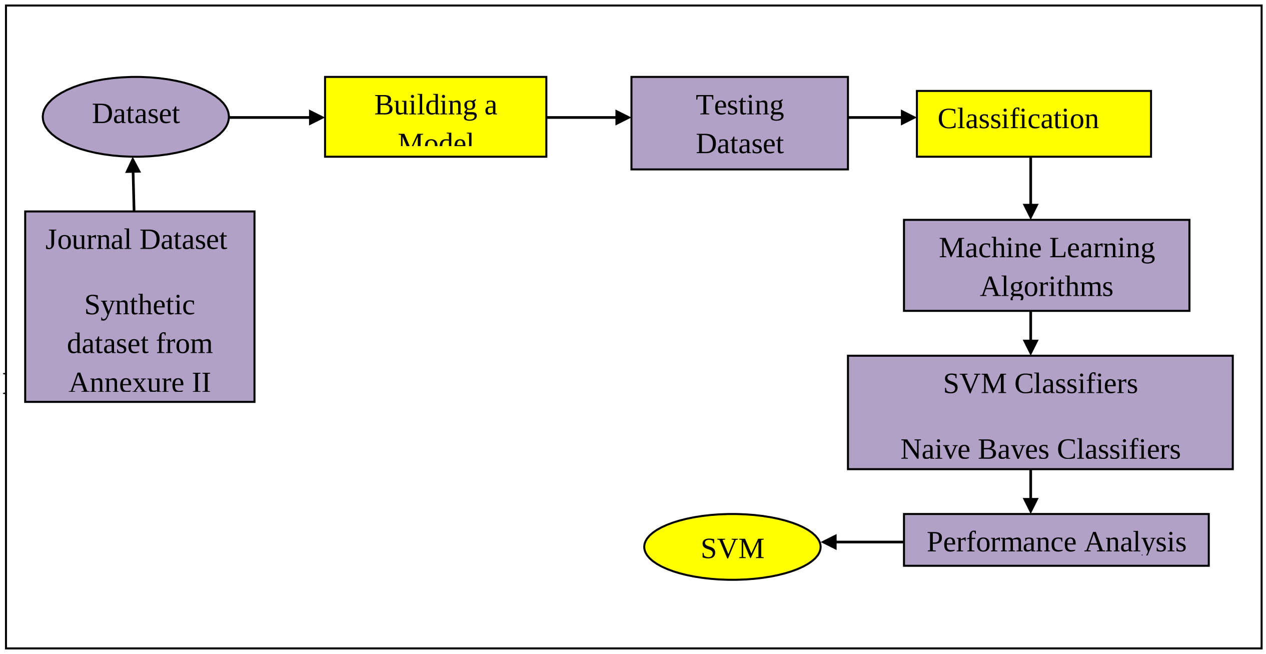
Shweta C. Dharmadhikari et.al [11] observed that proper classification of text documents requires information retrieval, machine learning and natural language processing (NLP) techniques. They aimed to focus on important approaches for automatic text classification which is based on machine learning techniques viz. Supervised, unsupervised and semi supervised. The research paper presented a review of various text classification approaches under the machine learning paradigm.

Shalini Puri et.al [10]proposed a new Fuzzy Similarity Based Concept Mining Model (FSCMM) is to classify a set of text documents into pre - defined Category Groups (CG) by providing them training and preparing on the sentence, document and integrated corpora levels along with feature reduction, ambiguity removal on each level to achieve high system performance. Fuzzy Feature Category Similarity Analyzer (FFCSA) is used to analyze each extracted feature of Integrated Corpora Feature Vector (ICFV) with the corresponding categories or classes. This model uses Support Vector Machine Classifier (SVMC) to classify correctly the training data patterns into two groups; i. e., + 1 and – 1, thereby producing accurate and correct results. The proposed model works efficiently and effectively with great performance and high - accuracy results.

Durga Bhavani Dasari et.al [12], observed that many documents are available in digital forms which needs text classification. For solving this major problem, present researchers focused on machine learning techniques: a general inductive process automatically builds a classifier by learning, from a set of pre classified documents, the characteristics of the categories. The main benefit of the present approach is the manual definition of a classifier by domain experts were effectiveness, less use of expert work and straightforward portability to different domains are possible. The paper examined the main approaches to text categorization which compared the machine learning paradigm and presented state of the art. Various issues pertaining to three different text similarity problems, namely, semantic, conceptual and contextual are also discussed.

Samuel Danso et.al [13] created a report on a comparative study of the processes involved in Text Classification applied to classifying Cause of Death: feature value representation; machine learning classification algorithms; and feature reduction strategies in order to identify the suitable approaches applicable to the classification of Verbal Autopsy text. They demonstrated that the normalized term frequency and the standard TFiDF achieved comparable performance across a number of classifiers. The results also proved Support Vector Machine is superior to other classification algorithms employed in this research. Finally, the authors demonstrated the effectiveness of employing a ’locally-semi supervised’ feature reduction strategy in order to increase performance accuracy.

III. METHODOLOGY



The synthetic journal dataset is created by extracting the information from the Annexure II journal list which is obtained from www.annauniv.edu/research website. It represents the journal list from various disciplines. This data set consists of four attributes, namely journal id, source title, ISSN number and country. The source title refers to the various journal's name, the ISSN number refers to the International Standard Serial Number of a particular journal paper and the country refers to the place where the journal is being published.

Dataset Pre-processing

Pre-processing is one of the important steps in text mining. This step is crucial in determining the quality in the classification stage [2]. It is important to select the significant keywords that carry the meaning, and discard the words that do not contribute to distinguishing between the documents [1]. In order to use these data they should be in a proper consistent form. So, we have used the pre-process methods to make the data legitimate. The pre-processing techniques used in this research work are Stop word removal and Stemming process. In this work the source title attribute from the dataset is subjected to these pre-processing methods.

Machine Learning

The use of machine learning has spread rapidly throughout computer science and beyond. Machine learning systems automatically learn programming from data. It is a set of methods that can mechanically detect patterns in data, and then uses the uncovered patterns to predict future data [5]. The real time applications of machine learning are Web search, spam filters, recommender systems, ad placement, credit scoring, fraud detection, stock trading and drug design. Machine learning is usually divided into two main types, namely [4]:

(i) **Predictive approach:** The goal of this approach is to learn a mapping from inputs x to outputs y, given a labelled set of input-output pairs $D = \{(a_i, b_i)\}N_i=1$. Here D is called the training set, and N is the number of training examples.

(ii) **Descriptive approach:** The second type of machine learning is the descriptive or an unsupervised learning approach. Here we are only given inputs, $D = \{a_i\}N_i =1$, and the goal is to find “attractive patterns” in the data. It is also called knowledge discovery.

A classifier is a system that inputs (typically) a vector of discrete and/or continuous feature values and outputs a single discrete value, the class [5]. For example, a bank manager analyses the bank loan application whether giving loan to a particular person is “risk” or “safe” based on his income, and its input may be a Boolean vector x = (x_1, . . . , x_j , . . . , xd), where xj = 1 if the jth word in the dictionary appears in the income and xj = 0 otherwise. A learner inputs a training set of examples (xi, yi), where xi = (xi,1, . . . , xi,d) is an observed input and yi is the corresponding output, and outputs a classifier. The test of the learner is whether this classifier produces the correct output y_t for future examples x_t. The machine learning algorithms used in this research work are SVM and Naïve Bayes algorithm.

SVM Classifiers

In recent decades the Support Vector Machines became one of the most popular machine learning algorithms for classification technique. It is a method, a special kind of a rule that produces classifiers with good predictive performance. It is explicitly based on a theoretical model of learning [6]. In Support Vector classification, the separating function can be expressed as a linear combination of kernels associated with the Support Vectors as,

$$f(x) = \sum_{x_j \in S} \alpha_j y_j K(x_j, x) + b$$

Where x_i refers to the training patterns and S denotes the set of support vectors.

The wide variety applications of SVM algorithms are text classification, facial expression and gene analysis etc. The advantages of the SVM classifiers are that they produce very accurate classification results and less outlier. Outliers in this research work refer to source title S_i which never falls into any one of the classes. The disadvantages of SVM classifiers are it is a binary classifier. To do a multi-class classification, pair-wise classifications can be used and they are computationally expensive [7].

Support vectors are the critical elements of the training set. The support vectors are the data points that lie close to the decision line or the optimal hyperplane. The elements of the training set that would change the position of the dividing hyper plane if removed.

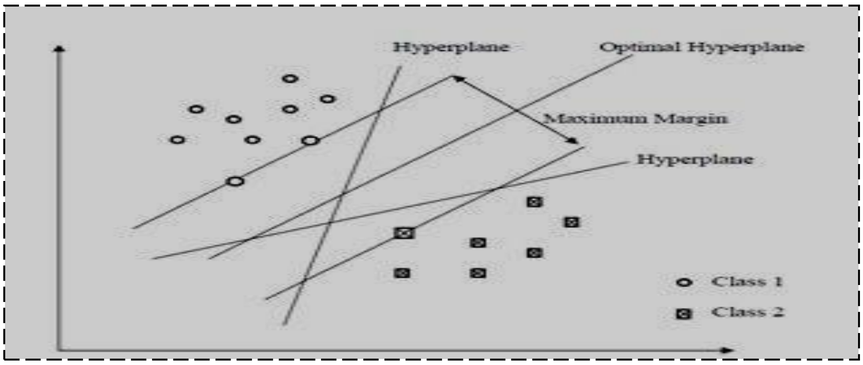


Figure 2: Optimal hyperplane for SVM classifiers

Training dataset using SVM classifiers:

Create input/output sets X , Y training set $(x_1, y_1).....(x_m, y_m)$. We have want to learn a classifier: y = f (x, α), where α are the parameters of the function. For example, if we are choosing our model from the set of hyperplanes in R_n, then we have: $f(x\{w,b\}) = sign(w \cdot w + b)$.

There are two kinds of risk for handling the training and test data in the SVM classifiers namely the empricial risk and the true risk. The following equations depicts the risks.To learn f (x, α) by choosing a function that performs well on training data[9]:

$$R_{emp}(\alpha) = \frac{1}{m}\sum_{i=1}^{m} l(f(x_i, \alpha), y_i) \rightarrow (1)$$

Where eqn(1) refers to the training error, *l* refers zero-one loss function, $l\left(y,\hat{y}\right)=1$ if $y\neq\hat{y}$ and 0 otherwise . R_{enp} is called the emprícial risk.By following this procedure we can try to minimize the overall risk,

$$R(\alpha)=\int l(f(x,\alpha),y)dP(x,y)\rightarrow(2)$$

Where the eqn(2) refers to the testing error, $P(x,y)$ is the known joint distribution function of x and y.

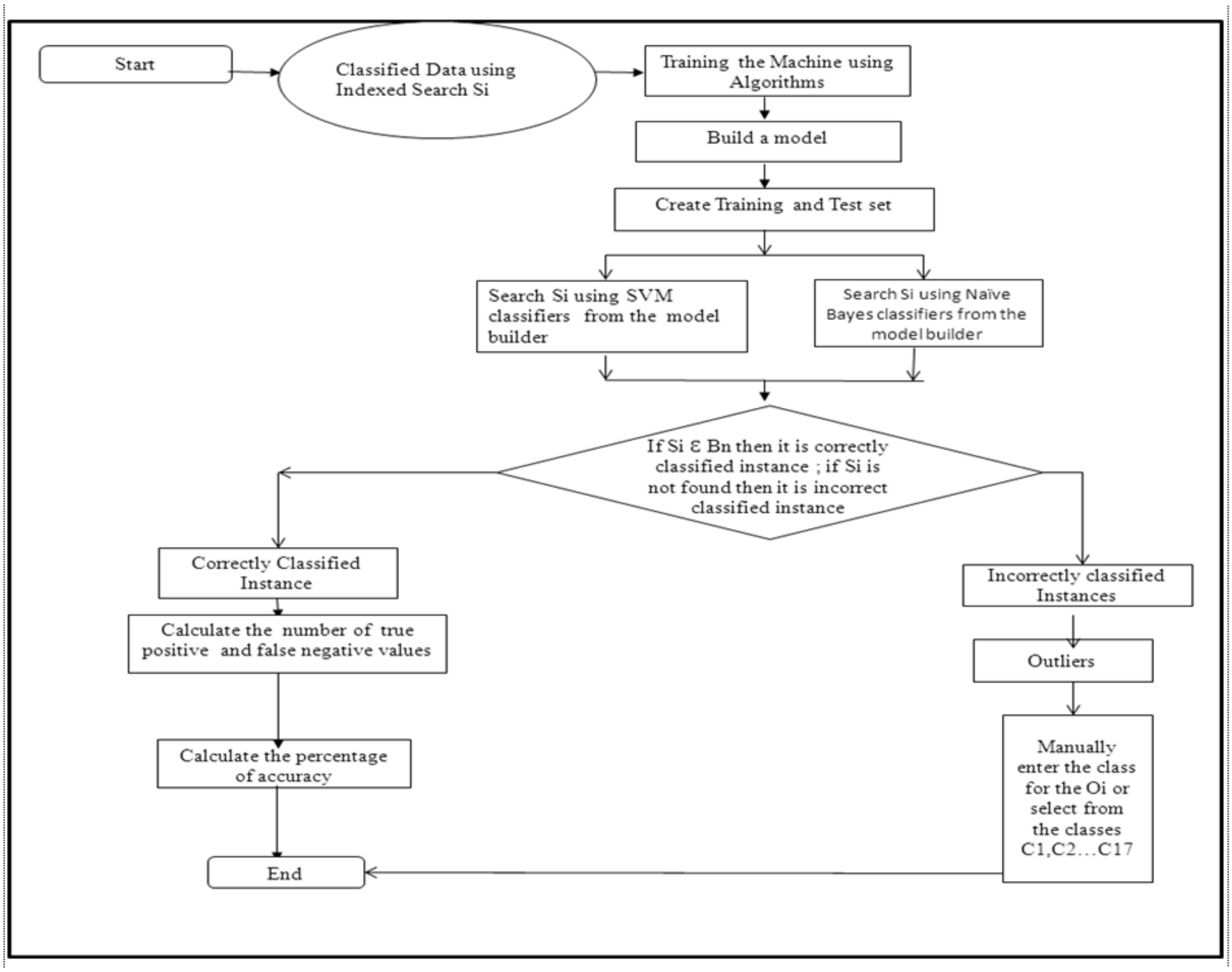
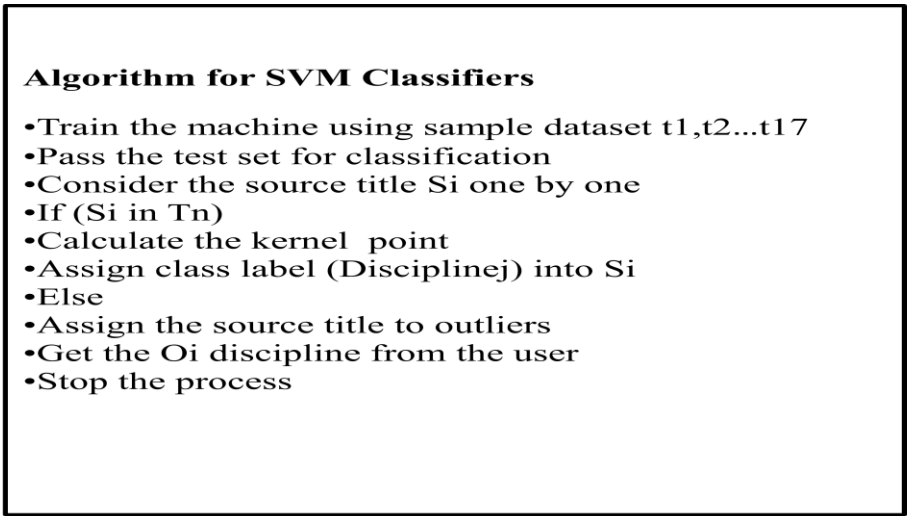

Start
Classified Data using Indexed Search Si
Training the Machine using Algorithms
Build a model
Create Training and Test set
Search Si using SVM classifiers from the model builder
Search Si using Naïve Bayes classifiers from the model builder
If Si Ɛ Bn then it is correctly classified instance ; if Si is not found then it is incorrect classified instance
Correctly Classified Instance
Calculate the number of true positive and false negative values
Calculate the percentage of accuracy
End
Incorrectly classified Instances
Outliers
Manually enter the class for the Oi or select from the classes C1,C2…C17

Figure 3: Flowchart for SVM Classifiers

After the classification task we find the closet pair of points first of all, we observe that, finding the closest pair of points in kernel space requires n^2 kernel computations where n represents the total number of data points. But, in case we use a distance preserving kernel like the exponential kernel the nearest neighbours in the feature space are the same as the nearest neighbours in the kernel space. Hence we need not perform any costly kernel evaluations for the initialization step [8].



Algorithm for SVM Classifiers

- Train the machine using sample dataset t1,t2...t17
- Pass the test set for classification
- Consider the source title Si one by one
- If (Si in Tn)
- Calculate the kernel point
- Assign class label (Disciplinej) into Si
- Else
- Assign the source title to outliers
- Get the Oi discipline from the user
- Stop the process

Figure 4: Algorithm for SVM classifiers

Naïve Bayes Classifiers

One of the simple classifier is the Naïve Bayes classifier, which makes independence assumption that for performing classification. A naïve Bayes classifier is a probabilistic classifier based on applying Bayes' theorem with strong independence assumptions. The structure of Naïve Bayes classifiers is depicted in Figure 5.

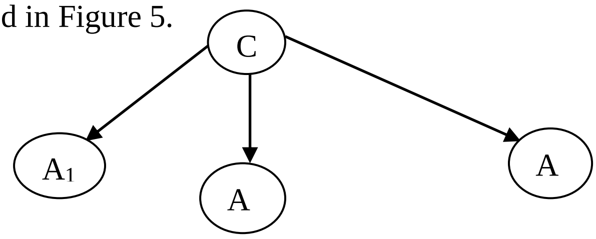


Figure 5: Structure of Naïve Bayes Classifier

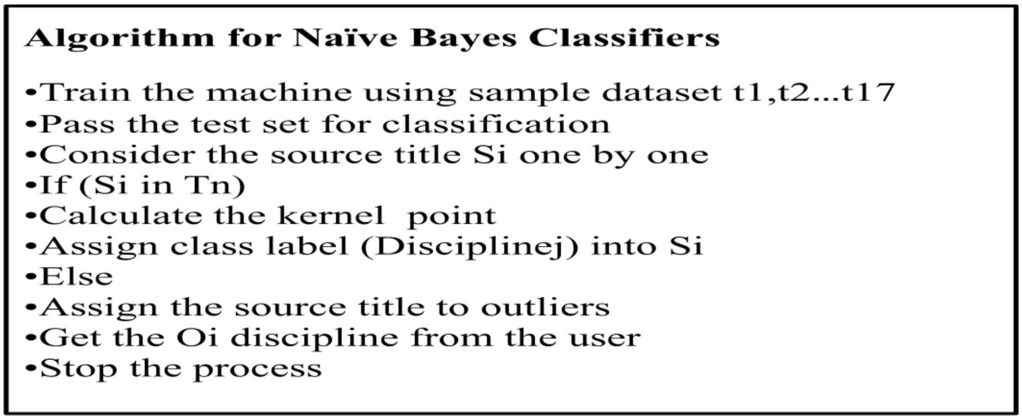
It shows the independence assumption among all features in a data instance. The Naïve Bayes Classifier assumes that the effect of each attribute on a class is statistically independent of all other attributes, this assumption, called class conditional independence. The advantages of Naïve bayes classifiers are that they be fastly trained, fast learning, simplicity and robust. The two best cases occur when they are completely independent and functionally dependent.

Methods classification using Naïve Bayes consists of the following steps: Input the dataset, pre-process the data set using the attribute source title, extract the keywords, create the training and test set for classification, use dictionary for identifying the terms and finally classify using the Naïve bayes machine learning algorithm.

Methods for Calculating Prior Probabilities and Conditional Probabilities of each class

The prior probability of each class is calculated by dividing the number of data instances with that class in the training set by the total number of instances in the training set[10].

The conditional probabilities for each feature value in the test data are calculated by getting the count of instances with that feature value in a particular class and dividing it by the count of instances with the same class in the training set. This is done for each class in the data set [10].The below given figure 6 illustrates the pseudo code for the training the dataset and perform the classification using the machine learning algorithm naïve bayes.



```
Algorithm for Naïve Bayes Classifiers

•Train the machine using sample dataset t1,t2...t17
•Pass the test set for classification
•Consider the source title Si one by one
•If (Si in Tn)
•Calculate the kernel  point
•Assign class label (Disciplinej) into Si
•Else
•Assign the source title to outliers
•Get the Oi discipline from the user
•Stop the process
```

Figure 6: Algorithm for Naïve Bayes Classifier

IV. EXPERIMENTAL RESULTS

Performance Measurements:

Performance measurement is generally defined as regular measurement of outcomes and results, which generates reliable data on the effectiveness and efficiency of programs. The performance measure in this research work is used to identify the best method for classifying the information. In order to measure the performance of the two method, four different criteria are used namely percentage of the correctly classified instances, percentage of the incorrectly classified instances, number of outliers and time taken.

Correctly classified instance

It is also called as true positive rate or the recall rate, which measures the proportion of actual positives which are correctly, identified instances.

True positive = correctly identified

$$TPR = \frac{TP}{P} = \frac{TP}{(TP + FN)}$$

Incorrectly classified Instances

It is the false positive rate or error in the predicted data for the test set.

False positive = incorrectly identified

$$SPC = \frac{TN}{N} = \frac{TN}{(FP + TN)}$$

$$FPR = \frac{FP}{N} = \frac{FP}{(FP + TN)} = 1 - SPC$$

Where **P** is positive instances and **N** is negative instances

Outliers

Outliers in this research work refers to source title S_i which never falls into any one of the classes arranging from $C_1, C_2, \ldots C_{17}$.Outliers is an examination used to find the number of S_i which are unlabelled after classification process. In order to handle the unlabelled instances the concept of outliers are used. These instances are stored in particular location S_x, from which the user can select the class for each instance S_i manually from the given classes C_n or he can also specify the class which he wishes.

Search time

The time complexity of an algorithm quantifies the amount of time taken by an algorithm to run as a function of the length of the string representing the input. Time complexity is commonly estimated by counting the number of elementary operations performed by the algorithm, where an elementary operation takes a fixed amount of time to perform. Thus the amount of time taken and the number of elementary operations performed by the algorithm differ by at most a constant factor. Search time measures the amount of required time for classifying the information.

Table 1: Classification Accuracy Measure in %

Algorithms	Testing set		
	Percentage of correct classified Instances (%)	Percentage of Incorrect classified Instances (%)	Percentage of outliers (%)
SVM algorithm	95.37	3.08	1.54
Naïve Bayes	90.10	4.88	5.01

The above table 5.3 depicts the percentage of the correctly classified instances, percentage of the incorrectly classified instances and the percentage of the outliers. By calculating their performance the accuracy for classification is obtained.

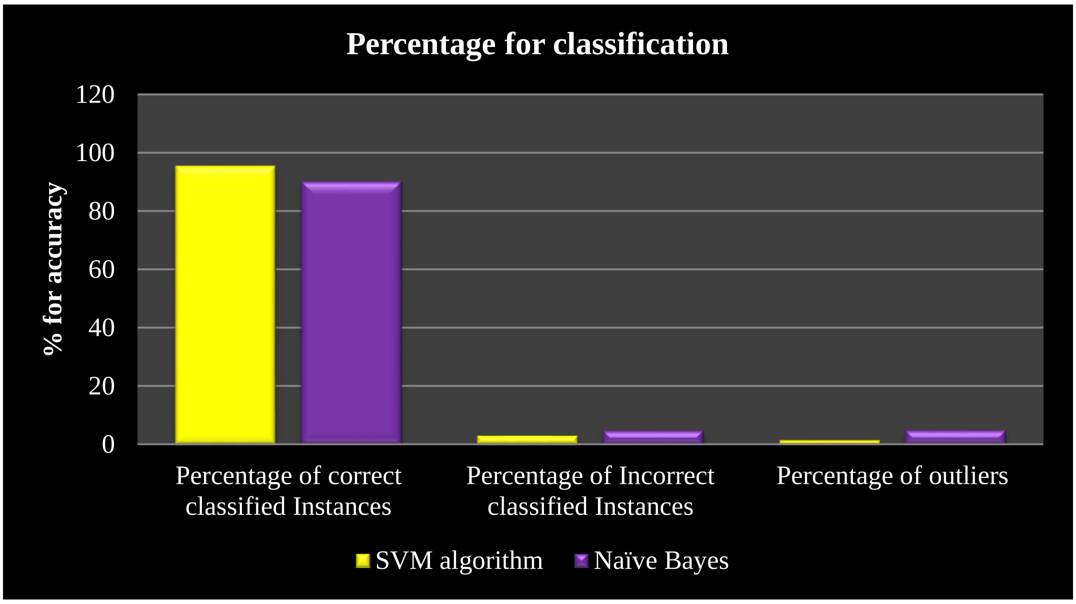


Figure 6: Percentage of Classification Accuracy

From the above given figure 7, it is analyzed that the SVM classifiers performs better than the Naïve Bayes algorithm. Therefore the SVM classification algorithm performs well because it attains lowest percentage of incorrectly classified and lowest percentage of outliers.

Time Taken

Table 2: Time taken for Classification

Method	Time taken
SVM classifiers	283 ms
Naïve Bayes classifiers	371 ms

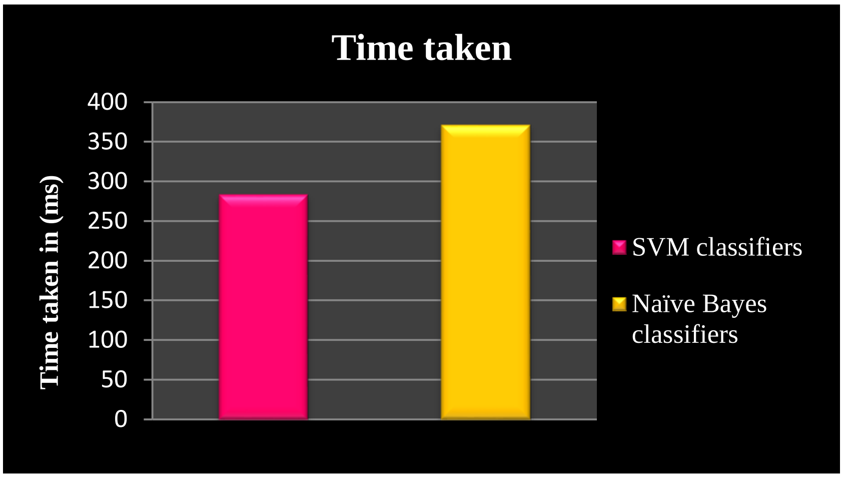


Figure 8: Time taken for classification

Figure 8 represents the execution time required for performing classification tasks by using SVM classifiers and Naïve bayes machine learning algorithm. From this result, it is known that the SVM classifiers requires minimum execution time compared to other algorithm.

REFERENCES

[1] Andreas Hotho, A Brief Survey of Text Mining, University of Kassel, May 13, 2005.

[2] Dmitriy Fradkin and Ilya Muchnik, Support Vector Machines for Classifcation, DIMACS Series in Discrete Mathematics and Theoretical Computer Science.

[3] Durga Bhavani Dasari & Dr . Venu Gopala Rao. K, Text Categorization and Machine Learning Methods: Current State of the Art, Global Journal of Computer Science and Technology Software & Data Engineering Volume 12 Issue 11 , Online ISSN: 0975-4172, 2012.

[4] http://www.cs.columbia.edu/~kathy/cs4701/documents/jason_svm_tutorial.pdf

[5] **http://eprints.pascal-network.org/archive/00000017/01/Tutorial_Marko.pdf,** Text Mining Tutorial.

[6] **http://www.tutorialspoint.com/data_mining/dm_tutorial.pdf,** Data Mining Tutorial.

[7] Ikonomakis M,Kotsiantis S, Tampakas V, Text Classification using Machine Learning Techniques, WSEAS transactions on Computers, Issue 8, Volume 4, August 2005, pp. 966-974.

[8] Kevin P. Murphy, Machine Learning,A Probabilistic Perspective, Cambridge, Massachusetts,London, England.

[9] Koby Crammer, Yoram Singer, On the Algorithmic Implementation of
Multiclass Kernel-based Vector Machines, Journal of Machine Learning Research , 2001.

[10] Pedro Domingos, A Few Useful Things to Know about Machine Learning, Department of Computer Science and Engineering, University of Washington, Seattle, U.S.A.

[11] Samuel Danso, Eric Atwell and Owen Johnson, A Comparative Study of Machine Learning Methods for Verbal Autopsy Text Classification, Language Research Group, School of Computing, University of Leeds, UK.

[12] Shweta C. Dharmadhikari, Maya Ingle , Parag Kulkarni, Empirical Studies on Machine Learning Based Text Classification Algorithms, Advanced Computing: An International Journal (ACIJ), Vol.2, No.6, November 2011

[13] Vishwanathan, M. Narasimha Murty, SSVM : A Simple SVM Algorithm, Dept. of Comp. Sci. and Automation, Indian Institute of Science, Bangalore, India.