An Improved Text classification for Unstructured Document

Naga Sudha D[1], Y Madhavee Latha [2]

[1]Research Scholar JNTUH, Hyderabad [2]Prof.,MRECW ,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, Shillloute Coefficient .

I.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.

II.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.

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.

III. EXISTING METHODS

For text classification supervised learning algorithms like Naive bayes,Support Vector Machine and Logistic Regression with labelled data is performed.In order to improve the accuracy of the classifier the feature representation and optimal parameter k is proposed.

IV. 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 %. TFIDF , 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 from daily haunt, cleaned, and properly organized, the terms (features) are identified, and a vector space representation created.

Pre-processing: The text document is represented in a word format in the preprocessing step. The preprocessing step inludes tokenization and stop word removal. The contents of the file is tokenized into
individual words.
Stop Word Removal:Common words like articles, preposition's and pronouns etc are called stop words and they are removed.

```
# Source: https://github.com/Xangis/extra-stopwords (MIT License) Spacy
STOPWORDS = ['అందరూ', 'అందుబాటులో', 'అడగండి', 'అడగడం', 'అడ్డంగా', 'అనుగుణంగా', 'అనుమతించు', 'అనుమతిస్తుంది', 'అయితే'
,
             'ఇప్పటికే', 'ఉన్నారు', 'ఎక్కడైనా', 'ఎప్పుడు', 'ఎవరైనా', 'ఎవరో', 'ఏ', 'ఏదైనా', 'ఏమైనప్పటికి', 'ఒక', 'ఒకరు', 'క
నిపిస్తాయి',
             'కాదు', 'కూడా', 'గా', 'గురించి', 'చుట్టూ', 'చేయగలిగింది', 'తగిన', 'తర్వాత', 'దాదాపు', 'దూరంగా', 'నిజంగా', 'పై', 'ప్ర
కారం',
             'ప్రక్కన', 'మధ్య', 'మరియు', 'మరొక', 'మళ్ళీ', 'మాత్రమే', 'మెచ్చుకో', 'వద్ద', 'వెంట', 'వేరుగా', 'వ్యతిరేకంగా',
'సంబంధం']
```

Feature extraction and selection: Feature representation is most important tasks of document classification. In feature representation of documents, documents are converted into feature vectors.
1.TFIDF Vectorization Model

2.Count Vectorization Model

Count Vectorizer: In binary vectorizer all the words in the vocabulary, the words that occur in the document at least once, is counted positive(1) where as the words that do not occur is not counted (0).

TFIDF Vectorizer: Count Vectorizer considers the frequency of words occurring in a document, it does it irrespective of how rare or common the word is. To over come this limitation TFIDF Vectorizer is used. It considers the inverse document frequency along with frequency of each word occurring in a document, in forming the feature vector.

To verify the clusters and reduce the curse of dimensionality is resolved for the most popular methods for dimensionality reduction. Principal Component Analysis is one which is used for Feature Extraction of telugu documents.

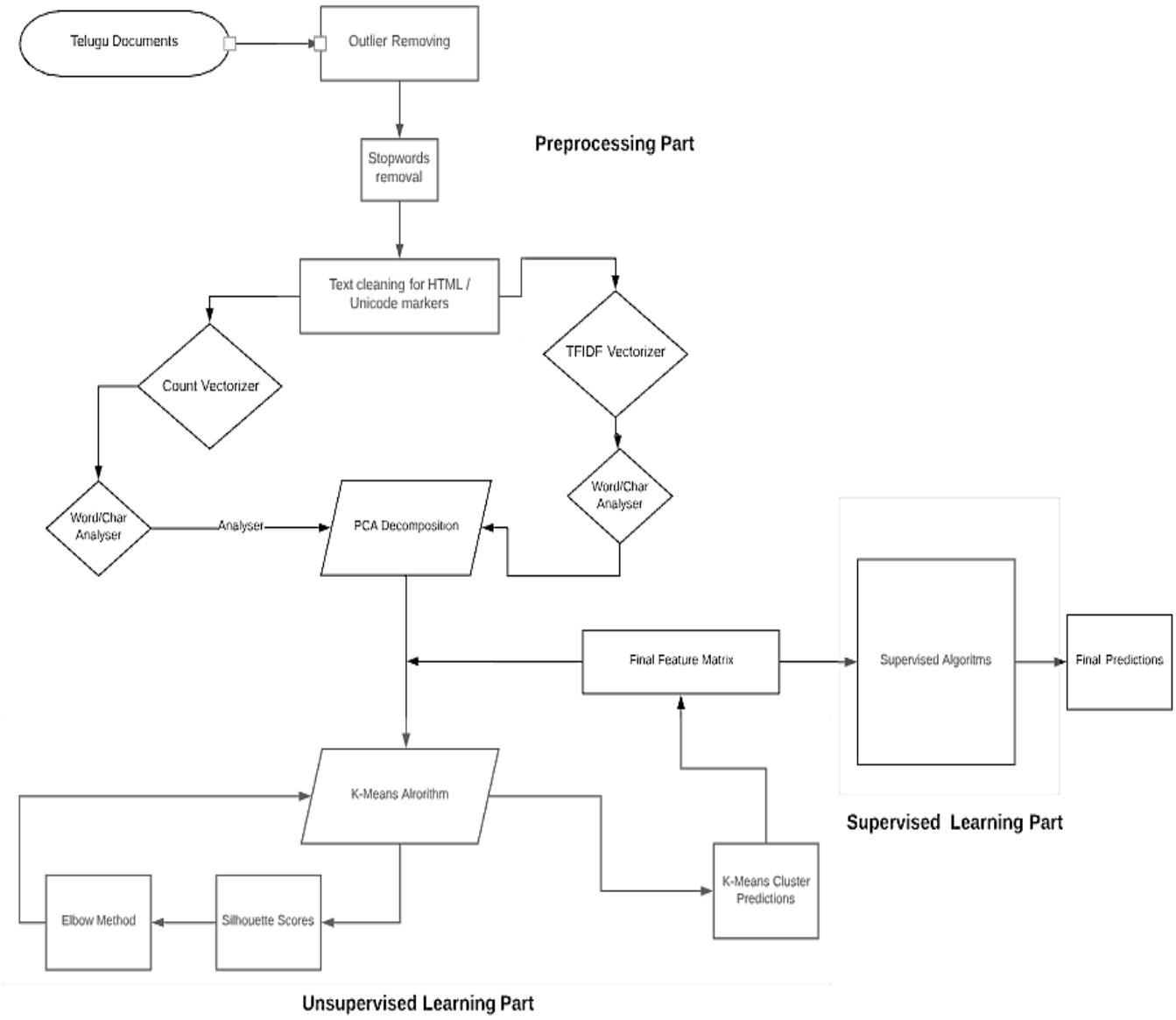

Fig 1.Semi supervised Text classification

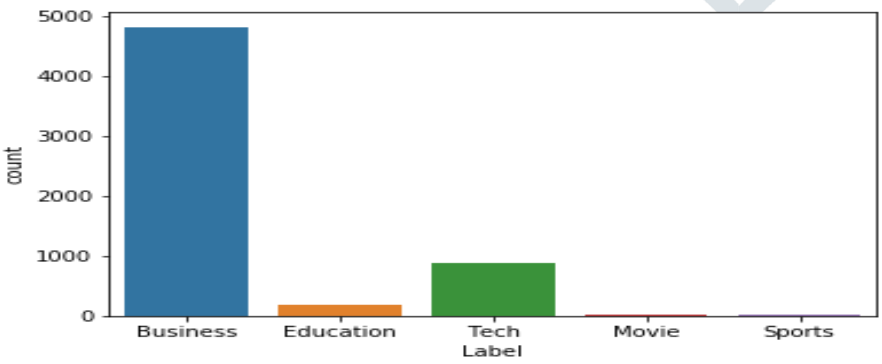

Fig 2. News Corpus

In this PCA as feature matrix reduction for unsupervised clustering to predict the cluster of the document while using t-SNE to visualize the cluster after reducing to 3 major components. The visualization of the clusters was done using plotly library. The following figure is the TSNE visualization using char analyzer using TF IDF vectorization method.

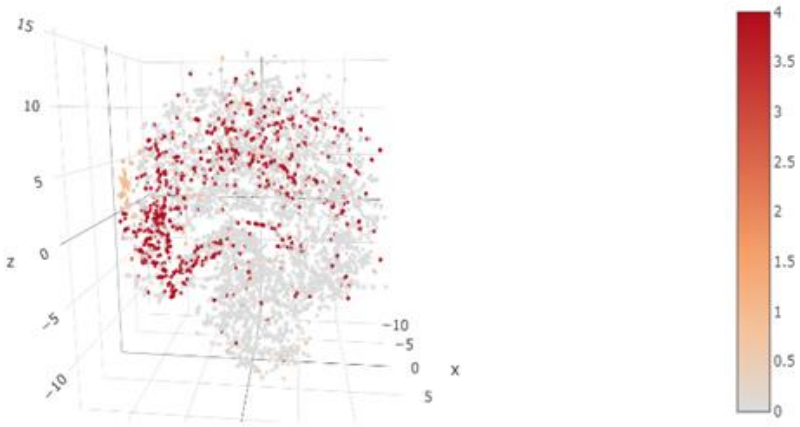

Fig 3. TSNE 3D Char Analyser after dimensionality reduction

Determining the optimal number of clusters in a data set is a fundamental issue in partitioning clustering, such as k-means clustering which requires the user to specify the number of clusters k to be generated. The optimal number of clusters is somehow subjective and depends on the method used for measuring similarities and the parameters used for partitioning. In direct methods consists of optimizing a criterion, such as the within cluster sums of squares or the average silhouette. The corresponding methods are named elbow and silhouette methods, respectively. Elbow Method Determining the proper number of cluster is one of the basic drawback in k-means algorithm. The average silhouette of the data is another lucrative and precise way for determining the natural number of clusters.

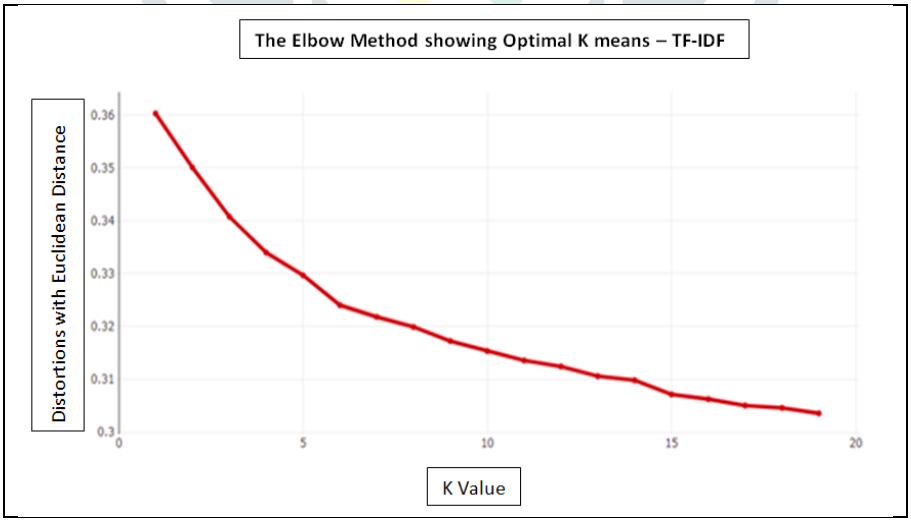

Fig 4. Elbow method with optimal K means –TF-IDF

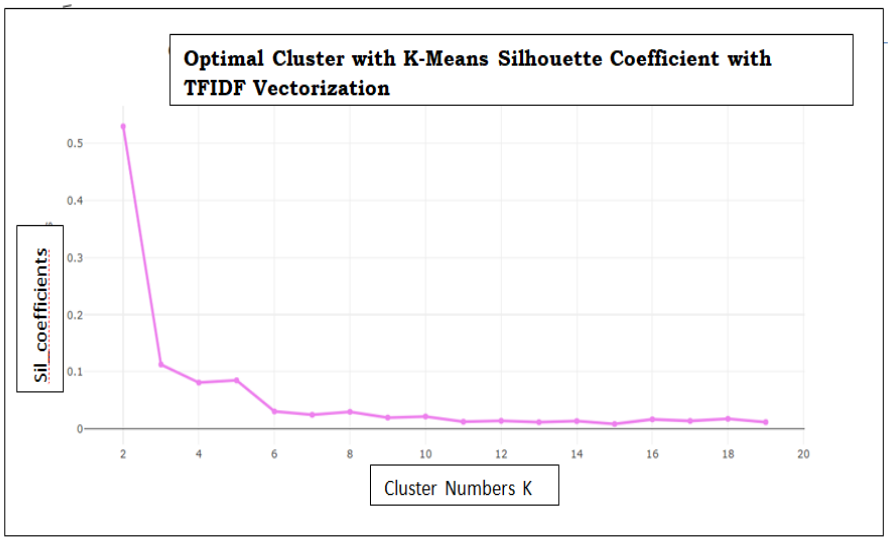

Fig 5. Optimal cluster with K-means silhouette Coefficient with TF-IDF

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.

I) 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.

II)Naive Bayes Classification Method: Naive 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.

$$P\left(\frac{Ci}{D}\right) = \frac{P(Ci)P\left(\frac{D}{Ci}\right)}{P(D)}$$

$$P\left(\frac{D}{Ci}\right) = \prod_{j=1}^{n} P(dj|ci)$$

Where P(Ci)= P(C=ci)=Ni/N

and P(dj|ci)= $\frac{1+Np}{M+\sum_{i=1}^{M} NM}$

III)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

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

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

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

Unsupervised optimal k feature is given to supervised algorithms like SVM, Logistic regression and Naive Bayes. By running the above three algorithms with various hyper-parameters ie. N-gram, threshold (min,max) values out of three algorithms SVM accuracy is increased for telugu text documents.

V.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 93.64

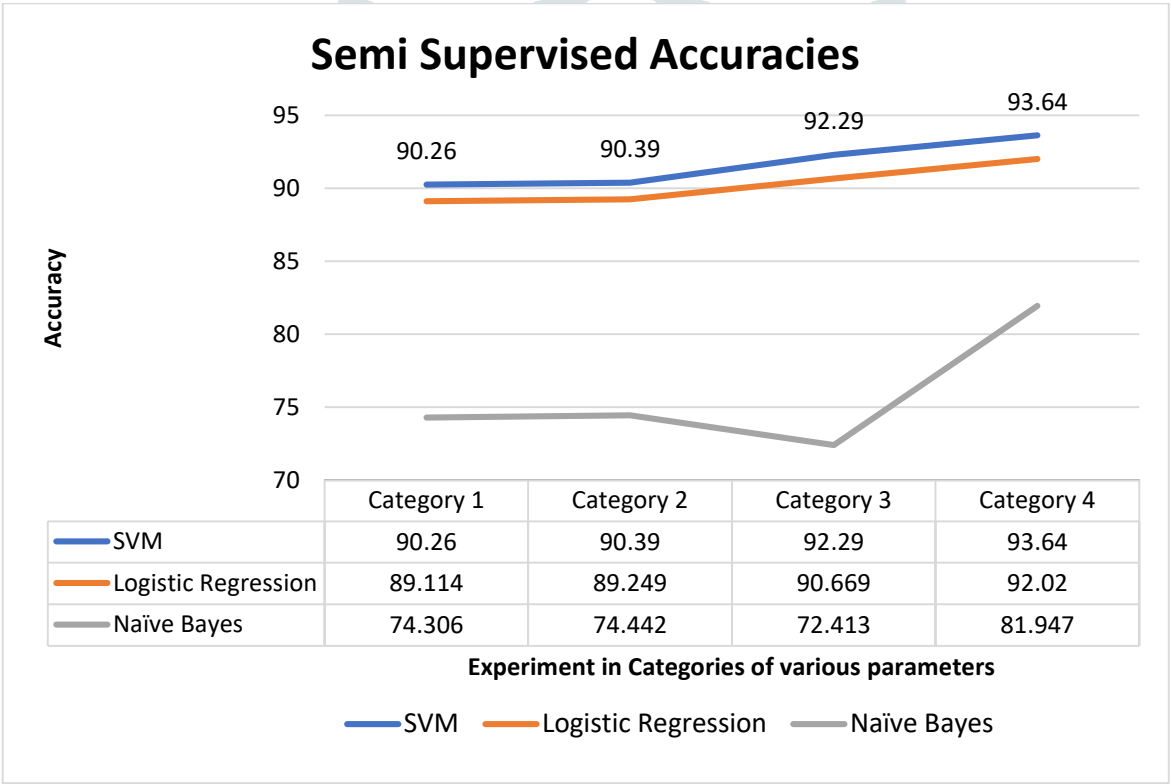

Fig5.Semi supervised classification accuracy of news articles

	Logistic Regression			SVM			Naive Bayes		
Feature	P	R	F1	P	R	F1	P	R	F1
Char N-gram 1-2	0.86	0.84	0.85	0.96	0.92	0.94	0.82	0.78	0.80
Char N-gram 1-3	0.92	0.89	0.90	0.96	0.93	0.94	0.84	0.82	0.83
Char N-gram 1-4	0.92	0.90	0.91	0.95	0.93	0.94	0.84	0.82	0.83
Word N-gram 1-1	0.82	0.80	0.81	0.88	0.87	0.88	0.78	0.74	0.76
Word N-gram 1-2	0.78	0.80	0.79	0.84	0.88	0.86	0.76	0.73	0.74
Word N-gram 1-3	0.78	0.74	0.76	0.82	0.84	0.84	0.71	0.73	0.72

Fig6. Precision,Recall and F-Score Values

VI.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.

VII.REFERENCES

[1]Mita K. Dalal, Mukesh A. Zaveri "Automatic Text Classification: A Technical Review", International Journal of Computer Applications (0975 – 8887),Volu me 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,(2016).

[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 Naive Bayes a Good Classifier for Classification?",International Journal of Software Engineering and Its Applications, Vol. 5, No. 3, July, (2011)