A Result On: Text Categorization Using Support Vector Machine

Aditya Jadhav[1], Pratik Patil[2], Shubham Patel[3], Abhishek Halkude[4], Prof.Nilufar Zaman[5]
[1, 2, 3, 4, 5] Dept of Computer Engineering
[1, 2, 3, 4, 5] Modern Educational Society's College Of Engineering, Pune

Abstract- *With the increase in use of Internet and development of Digital Technology, data has increased by many folds. Information in raw or unorganized form (such as alphabets, numbers or symbols) that refer to, or represent, conditions, ideas, or objects. Data is limitless and present everywhere in the universe. Data may consist images, audios, torrents, e-books, articles, textual content, videos etc. There is a need for an efficient tool to automatically classify the data into various categories. Textual Content is one of the most used data on Internet, by people and this data has significantly increased in last few years. This textual data needs to be categorized or labelled for convenience of organizing, searching and maintaining on the Internet. By categorization we can easily search, extract and store the large amount of resources in a well-organized way. Automated text categorization is a process that assigns pre-defined category labels to new documents based on the contents. Our primary goal is to develop a Text Categorization model to classify the text into pre-defined classes using Support Vector Machines.*

Keywords- Supervised Learning, K-Features, Feature Selection, Text Classification.

I. INTRODUCTION

With the wide use of Internet the data on the web has seen an exponential rise. Data of all sort and types is getting accumulated day-by-day.1,209,600 new data producing social media users each day. Large amount of data remains unstructured and uncategorized even today. To categorize Textual Document we use Text Categorization. Text Categorization is the process of classifying information in text format by its content. Text databases consist of large collections of documents from various sources such as news articles, research papers, digital libraries, e-mail messages and Web pages. Text Categorization, which is also referred to as Text Classification, is the task of automatically sorting a set of documents into categories from a predefined set, Text Categorization comes under the broader domain of text mining, which is the general term for the process of deriving any information from a given text using a variety of text processing techniques. Text classification with the help of SVMs is a supervised learning.

In text classification, a labelled dataset (training set) is provided to the classifier for training the classifier i.e. supervised learning. SVM is a supervised-learning algorithm. It means we will need provide manually some labelled dataset. Then the SVM is trained using this labelled dataset. After training the classifier it is possible to label the unlabelled document. Automatic text classification has always been an important application since the inception of digital documents. Text Categorization is now being used for various applications, for example: it is possible to classify web pages into different categories to speed up the Internet search, which is very useful for some search engines like Yahoo and Google, Email spam filtering,Document sorting, Expert recommendation systems, Population of hierarchical catalogues of Web Resources etc. The advantages of Automated TC is that the approach used here are that the accuracy achieved is comparable to that done with the help of human expertise but the savings in labour work and expert and extensive knowledge base as well as reduction in human errors. A common approach used for feature reduction is selecting features. Feature Selection is the process of selecting only the relevant features for the categorization. The main approaches of feature selection are the filter approach and wrapper approach[13]. Most filter approaches calculate class dependent feature scores. Using combination operation may bias the the feature importance for discrimination [14].

The rest of the paper is organized as follows: In Section 2, we introduce related work and previous work on document representation, SVC, and feature selection techniques for text categorization. In Section 3, we present our proposed Bayesian method for details. Experimental results are described and analyzed in Section 5, and a conclusion is given in Section 6.

II. RELATED WORK

F. Sebastiani [1] has defined text categorization as "the task of automatically sorting a set of documents into categories or classes from a predefined set" and has stated that it was in the domain of both information retrieval and machine learning. As such, several techniques derived from these domains have found application in implementing text classifiers.

Thorsten Joachims [2] proposed the use of Support Vector Machines for text classification and had also demonstrated that SVMs could offer better performance for text classifiers as compared to other well-known machine learning techniques such as Naive Bayes and K-Means classifiers.

N.Cristianini, [3] had made a detailed discussion on the working of SVMs. A popular class of text classifiers in the Naïve Bayes classifiers based on the Bayes' theorem. In this regard, Leung defined Bayesian classifiers as statistical classifiers that can predict the probability of a particular sample belonging to a particular class [4]. A variant of Naïve Bayes called Multinomial Naïve Bayes (MNB) is also often used in solving text categorization problems as evidenced in the work by Frank and Bouckaert [5] who had proposed a method that improves the efficiency of MNB classifiers used for text categorization by improving the performance of MNB in the context of unbalanced datasets.

Toker and Kirmemis, [6] developed document organising application which implemented the k-NN based text classifier
Guo and Greer, [7] found their work comparable to SVM-based text classifier as they combine the strengths of Rocchio classifier and k-NN classifier.

III. PROPOSED METHODOLOGY

Text classification is a significant task in document processing and sorting. The goal of text classification is to classify a set of documents into a fixed number of predefined classes. A document may belong to single class, multiple class or no class. When classifying a document, a document is represented as a "bag of words". Text Categorization is a part of Information Extraction but process of both are different from each other. Rather, a simple text classification task involves only the counts of words that appear in the document, from the count, the main topics of that the document are identified e.g. if in the document, football word comes frequently then "football" is assigned as its topic.

There are two phases in the Classification. They are Training Phase and Testing Phase.

A) *Training Phase*:

It is also called as Learning Phase; the set of documents used for *Learning Phase* is called training set. It describes a set of predetermined classes. Each document in the training set is assumed to belong to a predefined class [2][3].

B)*Testing Phase*:

In this step of classification unlabelled documents are classified. The known label of test document is compared with the classified result to estimate the accuracy of the classifier [8][9][10].

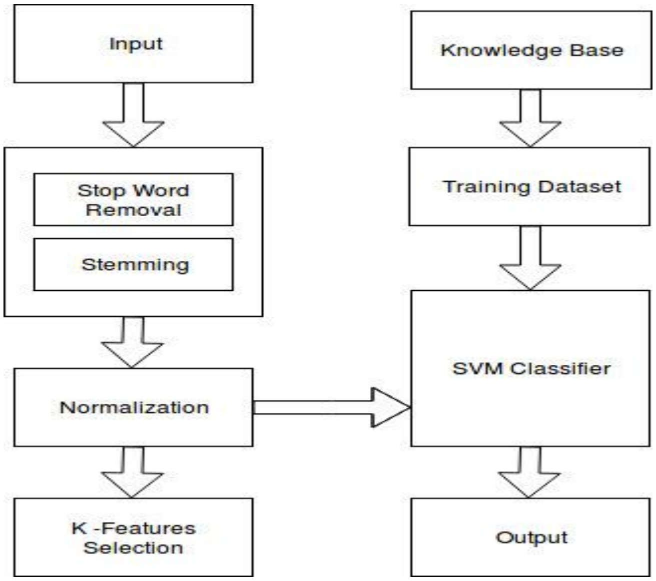

Fig. 1: Architecture of Text Classification

In the proposed architecture, an input in form of textual dataset is provided to pre-processor. The pre-processor uses stop word removal technique and then the remaining words are passed on for stemming. In STOP WORD REMOVAL, the words that are most common are eliminated (ex: and, or, of, the, etc.). During the stemming process, the ends of the words are chopped off to derive the root words (ex: playing - play).

In normalization, the undesirable characteristics of data are deleted to improve integrity & consistency of database. Normalization improves quality of database thus helping in faster classification. In k- feature selection best features will be selected based on ranking of occurrence's. In training phase a previously labelled document (training dataset) will be provided for training of the classifier. Finally the features will be given to the classifier for classification.

Using supervised learning algorithms [11], the main objective is to train classifiers from known datasets and perform the classification automatically on unknown datasets (unlabelled documents). Figure 1 shows a block diagram of Architecture of Text Classification. The main advantage of the system is it's performance and efficiency. This is achieved by multiple pre-processing techniques that help eliminate redundancy and independent terms.

Support vector machine: Structural Risk minimization is the main principle on which SVMs are based. The main idea of structural risk minimization is to find a hypothesis h for which can guarantee the lowest true error [2]. SVM is an universal learner algorithm i.e. they can learn linear threshold function. SVMs have a remarkable property that its learning can be independent of the dimensionality of the feature space. The complexity of hypotheses is measured based on the margin with which they separate the data, not the number of features. If the data is separable with a wide margin using functions from the hypothesis space, it is the possible to generalize even in the presence of very many features.

IV. PARAMETERS FOR TEXT CATEGORIZATION

In statistical analysis of binary classification, the F_1 score (also F-score or F-measure) is a measure of a test's accuracy. It considers both the precision *p* and the recall *r* of the test to compute the score: *p* is the number of correct positive results divided by the number of all positive results returned by the classifier, and *r* is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive). The F_1score is the harmonic average of the precision and recall, where an F_1 score reaches its best value at 1 (perfect precision and recall) and worst at 0.

$$F_1 = 2 * \left(\frac{precision * recall}{precision + recall}\right)$$

Precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances. In simple,Precision is the percentage of retrieved documents that are in fact relevant to the query.

Precision = |{Releavant}∩{Retrieved}| / |{Retrieved}|

Recall (also known as sensitivity) is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. Recall is the percentage of documents that are relevant to the query and were in fact retrieved.

Recall = |{Relevant}∩{Retrieved}| / |{Relevant}|

ROC curves are frequently used to show in graphical way the connection/trade-off between clinical sensitivity and specificity for every possible cut-off for test or a combination of tests. In addition the area under the ROC curve gives an idea about the benefit of using the tests.

V. RESULT

In our experiment we compared linear support vector classifer with Multinomial Naïve Bayes using class specific feature selection method. It has been seen that the classification increases slightly using SVM, the comparison result has been shown in following figures using various parameters.

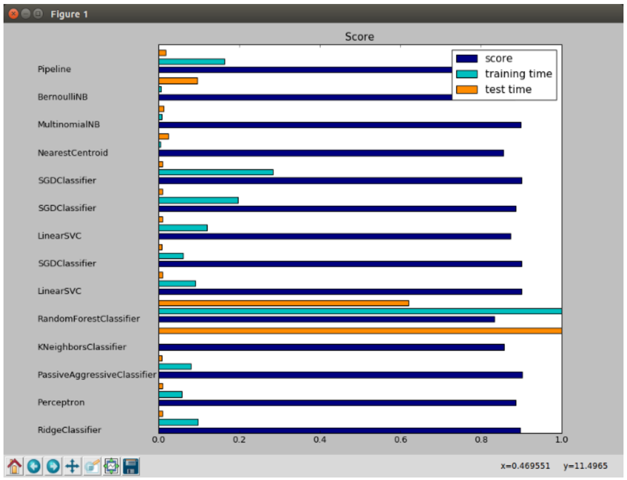

Fig 1: F_1 Score

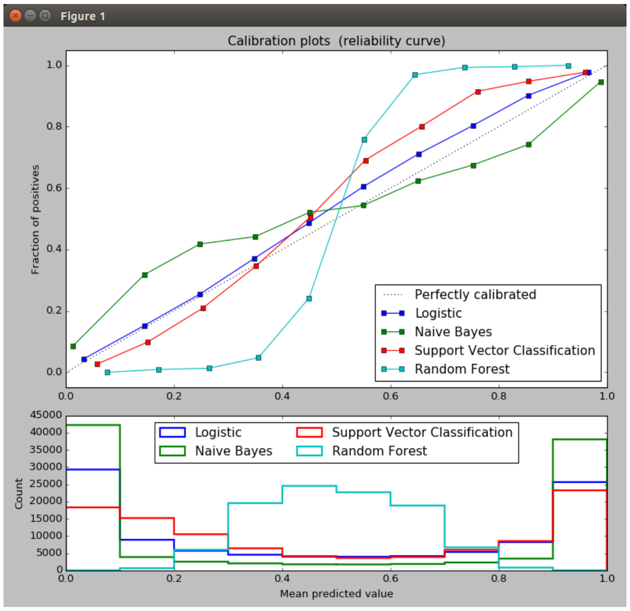

Fig 2: Calibration plots

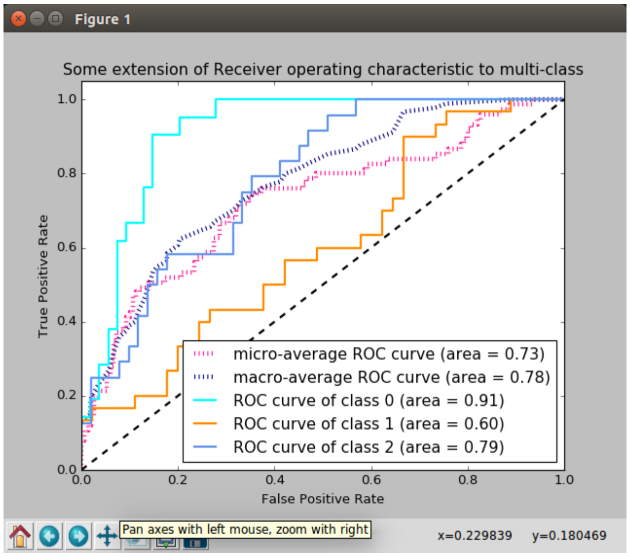

Fig 3: ROC curve

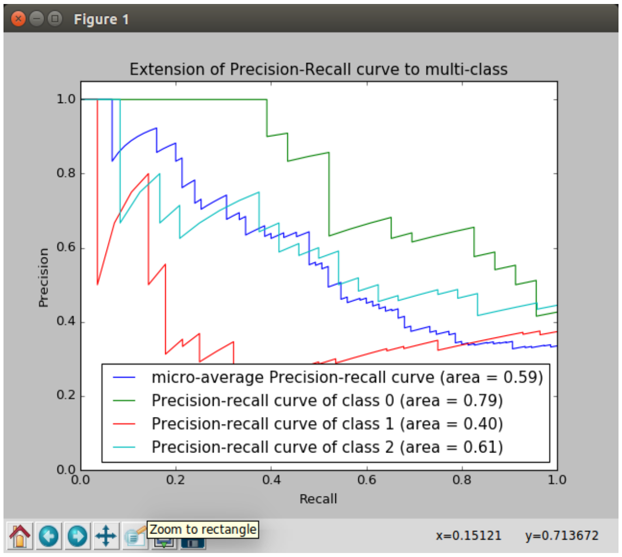

Fig 4: Precision-Recall curve

VI. CONCLUSION

In this paper, we used a SVM classification approach for atomatic text classification using class specific feature. The experiments conducted on various data sets showed that SVM is more promising than Naïve Bayes. Text Categorization is mainly done to ease the efforts required in maintaining and organizing the data as well as for Information Extraction. Labelling a Document in to appropriate category is of utmost importance for accurate Data Mining. Thus our project focuses on labelling a text dataset accurately by using SVM classifier. Using SVM as classifier helps for efficient classification, more importantly accurate classification. Classifications reduces high dimensional space to low dimensional space by dimensionality reduction feature. Using NIPS 2015 papers dataset SVM algorithm may outperformed Naive Bayes classification algorithm on the assumption that terms used in document are independent. The Bayesian classification used with class specific feature showed slightly low performance as compared to SVM.

REFERENCES

[1] F.Sebastiani, “Text Categorization”, UK, pp. 109-129, 2005.

[2] T. Joachims, Text Categorization with Support Vector Machines : “Feature Selection Methods for Text Classification , Universitat Dortmund, LS VIII, 1997.

[3] R. R. Bouckaert, Naive Bayes classifer for text classification PKDD 2006.

[4] K. Ming Leung, Naive Bayesian Classifier, Risk Engineering, 2007.

[5] N. Cristianini, Support Vector and Kernel Machines, International Conference on Machine Learning, June 28, 2011.

[6] Toker and Kirmemis, Text Classification using k Nearest Neighbor Classification, Survey Paper, Technical University.

[7] G. Guo et al., Using kNN model for text categorization, Soft Computing 10.5, 2006:

[8] Gurpreet S. Lehal, August 2009 “A Survey of Text Mining Techniques and Applications”,Journal of Emerging Technologies in Web Intelligence, VOL.1, NO. 1.

[9] Jiawei Han, Michelin Kamber, 2001, “Data Mining Concepts and Techniques”, Morgan Kaufmann publishers, USA, 70-181

[10] Megha Gupta, Naveen Aggrawal, 19-20 March 2010, “Classification Techniques Analysis”, NCCI 2010 - National Conference on Computational Instrumentation CSIO Chandigarh,INDIA.

[11] HaiyingTu, Jianhui Luo and Krishna R. Pattipati (2005), “Experiments on Supervised Learning Algorithms for Text Categorization”, International Conference, IEEE computer society.

[12] S. Kay and H. He, “Toward optimal feature selection in naive Bayes for text categorization,” IEEE Transactions on Data Engineering, 2016.

[13] B. Tang “ENN: Extended nearest neighbor method for pattern recognition [research frontier],” IEEE Computational Intelligence Magazine, vol. 10, no. 3,2015.

[14] “Toward Integrating Feature Selection Algorithms for Classification and Clustering” by Huan Liu and Lei Yu.