Performance of Classifiers on Newsgroups using Specific Subset of Terms

Deepanshu

Abstract*: Text Classification plays a vital role in the world of data mining and same is true for the classification algorithms in text categorization. There are many techniques for text classification but this paper mainly focuses on these approaches Support vector machine (SVM), Naïve Bayes (NB), k-nearest neighbor (k-NN). This paper reveals results of the classifiers on mini-newsgroups data which consists of the classifies on mini-newsgroups data which consists a lot of documents and step by step tasks like a listing of files, preprocessing, the creation of terms(a specific subset of terms), using classifiers on specific subset of datasets. Finally, after the results and experiments over the dataset, it is concluded that SVM achieves good classification output corresponding to accuracy, precision, F-measure and recall but execution time is good for the k-NN approach.

***Keywords*: Classifiers, k-NN, NaïveBayes, Text Classification, SVM.**

I. INTRODUCTION

Collection of words makes a document which represents its meaning. A lot of forms of data which is created such as news outlets, social networks like facebook, twitter, records of patterns, facebook, twitter, records of patterns, health care insurance data etc and in modern era, the key resource is information and its labeling for which various methods[1] are there to pick out information from huge data and document classification(TC) task[2] is one of them. The labeling of documents which are in the electronic forms needs labeling methods to categorize these contents. Several learning approaches can be used for the TC task[3] as well as various general steps are included in it which are as follows:

1. Text Preprocessing

It is the main task of text categorization task[4].Unlike the traditional works which contain feature extraction[5], feature selection[6] and classification techniques[7], preprocessing generally consists of steps[8] like tokenization followed by filtering, morphological analysis of words and stemming.

Tokenization: Here the tokens are known as words and phrases, these are formed after breaking the sequence of character and at the same time ignore the punctuation marks[9].

Filtering: Filtering is used to remove the stop words which are some of the words in the documents that usually occurs like conjunctions, punctuations etc. These words can be removed [10] because they do not contain much content information.

Lemma Finding : To make the single item from various infected word forms i.e. the morphological analysis.

Stemming: 'Stem' means the root of the derived words/ phrases. Stemming algorithms[11].

2. Vector Space Model(VSM)

VSM is a way which is used to show documents into the numeric vectors or variable which have a numeric value shows the weight(value for its importance) of the word. This way is very efficient in the analysis of documents[12].

3. Classification

Text classification aims to label predefined classes to text documents. Classification is defined as follows: If we have a training set X={p1, p2…….,pn} of documents, such that each document pi is labeled with a label qi from the set Z= {q1, q2, …..qk}. The main work is to find classification model (classifier) f where f: X → Z, f(x) = Z(3) which can assign the best class marking to new document d (test instance). For full overview of different classification techniques see[13],[14]. Different classification approaches which are used in this paper are as follows:

3.1 Support Vector Machine(SVM) Approach

This algorithm[15] is supervised learning algorithm and main requirement of SVM is that it required both +ve and -ve training dataset, to find surface which decides/separates the positive and negative data in n^{th}-dimentional.[16] there are two main drawbacks of SVM which are as follows:

1. It is applicable only on the binary classification.
2. It is difficult to represent document into numerical vectors.

3.2 Naïve Bayes

It is based on the probability generally on the bayes theorem[17]. Only less amount of training dataset is required[18]. In other words in the field of data mining, the naïve bayes(NB) classifier belongs to the cluster of the probabilistic classifier which is related to apply Bayes theorem with naïve supposition. Under the framework of naïve bayes(probabilistic) there are two feature model: Binary valued feature model and real-value model of feature which are used in Bernoulli NB and the multinomial respectively.

3.3 k-Nearest neighbor(k-NN)

The general definition of KNN is that, the two points(k,1) distance in the plane with co-ordinates(abscissa, ordinate) k=(x,y) and l=(a,b) can be calculated as:

$(k,l)=d(l, k)= \sqrt{(x-a)^2 + (y-b)^2}$

Mainly used in pattern recognition, k-NN[19] is not based on parametric method.

Input to the k-NN: Input which is given to k-NN consists of k close examples of training in feature space.

Output(for classification): class membership, if k=1 it means object marked to the class of single closest neighbor.

II. RELATED WORK

Bo Tang et al. in [20]

This paper shows the Bayesian (probilistic) approach for TC using specific feature subset for each class. For the application of class-specific features for categorization they follows baggenstoss's PDF projection theorem(PPT) and make a rule related to Bayesian. In this paper due to class-specific features, it allows to get the most imperative features for distribution and they derived a naïve Bayes rule that follows the PPT different to previous one.

Dino Isa et al.[21]

This paper reveals about the hybrid classification method with the help of NB and SVM. For vectorization, the Bayes formula was used, which give the categories based on probability that the topic(input) can relate to that category. Predetermined categories(set of topics) were taken for instance those found in "20 newsgroups". The vectorization part of the technique is same for both the classifier and finally, SVM classifier is applied on VSM for final output. The noticeable thing is discussed in the paper for improvement of pure Naïve Bayes classification approach.

Aggrwal & Zhai(2012)[22]

In this mainly discussed about the specific changes that can be applied for text classification. The algorithms about which they elaborate are decision trees, rule(pattern)-based classifiers, SVM classifiers, Bayesian classifiers, k-nn, neural network classifiers etc. They focused on the feature selection in TC and described about the performance metrics which are related to each algorithm.

V.Vaithiyanathan et al.[23]

This paper described the performance of the various classifiers. They took three datasets(UCI) for the experiment. Classifiers performance can be affected by many factors which are datasets, no. of tuples and attributes, types of attributes, system configuration & finally, showed that the multilayer perceptron gave better performance.

III. PROPOSED WORK

The problem of classification is a supervised classification problem. In this work, we choose three newsgroups dataset which are comp.os.ms-windows.misc, comp.graphics, alt.atheism and classifiers on which we measure the classification accuracy, precision, f-measure, recall are SVM, Naïve Bayes, k-NN.

Experimental Setup: Matlab

The program is designed so as to split the text data into 2 sets-training & testing. The entire dataset is split to 60% as train data and 40% as test data.

Steps to apply classification approaches on dataset are highlighted as:

1. Firstly reading the list of files which are comp.os..ms-windows.misc, comp.graphics, alt.atheism.
2. Preprocessing
   - Tokenization
   - Filtering
   - Lemmatization
   - Stemming
3. Terms Finding
4. Vector Space Model(VSM)
5. Classification Approaches(SVM, Naïve Bayes, k-NN)
6. Performance Measures in form of F-measure, accuracy, recall, precision and time complexity.

In the third step, we took specific terms which are shown in fig 1.1

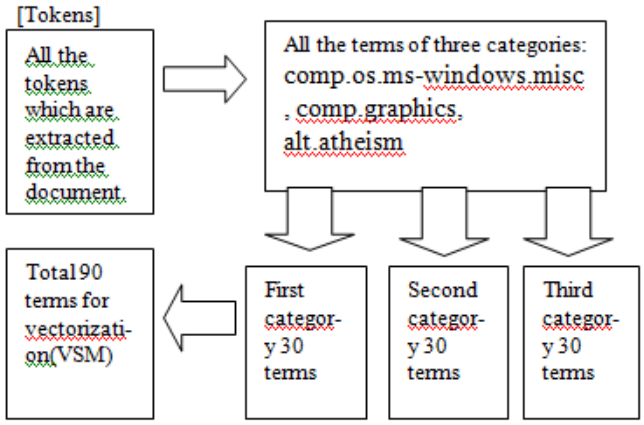

Fig 1.1 Specific terms are selected for VSM.

In this work firstly above mentioned four steps were performed and after that SVM, Naïve Bayes, the k-NN algorithm applied to the vector space model. After performing all the steps we measured the accuracy, f-score, recall and the execution time.

IV. EXPERIMENTAL RESULTS

On the basis of the experiment, it is clearly seen that the SVM outperformed on the other classifiers in accuracy, recall, precision, f-score and execution time for this classifier is also more than other classifiers. The order of the classifiers performance in terms of metrics which are shown in table 1.1 is as follows:

SVM > kNN > Naïve Bayes

But for the execution time the order is as follows:

Naïve Bayes > SVM > kNN

	accuracy	*precision*	*recall*	*f-mea.*	time
kNN	78.777	68.792	68.8	68.2	0.02
NB	76.111	74.592	63.2	58.2	0.33
SVM	83.333	74.909	75.2	74.7	0.27

Table 1.1 Comparison of three classifiers on the basis of accuracy, recall, F-measure, precision, execution time.

Mathematical forms:

a) Accuracy- No. of correct predictions/Total no. of predictions made

b) Precision- No. of true positives/n =No. of true positive and false positive.
c)Recall- No. of True positives/No. of true positives and the no. of false –ve.
d)F-measure/F-score=
2*((precision*recall)/(precision+recall)).
Graphical Representation of all the metrics corresponding to their classifier is in following figures.

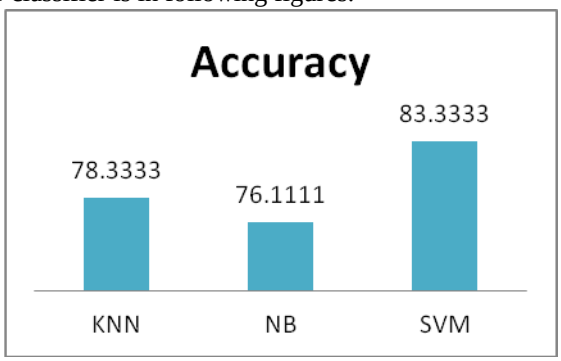

Fig.1.2 Accuracy of three classifier

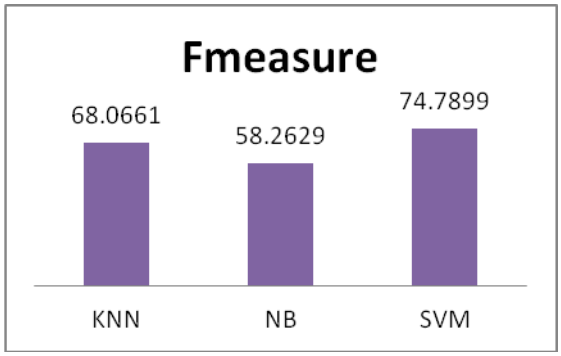

Fig.1.3 f-measure of three classifier

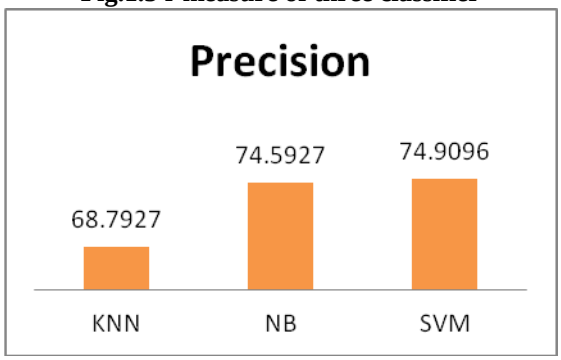

Fig.1.4 precision of three classifiers

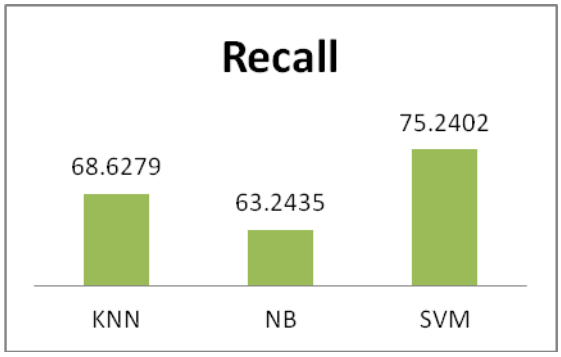

Fig.1.5 recall of three classifiers

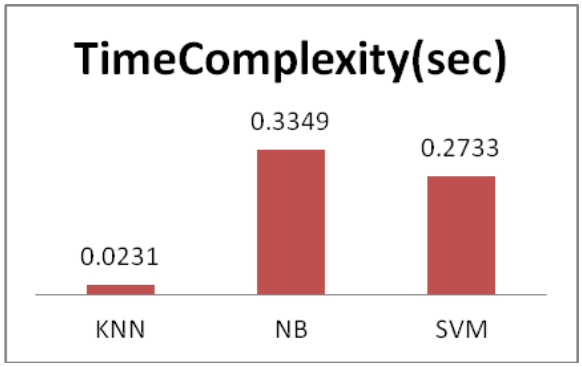

Fig.1.6 Execution time of three classifiers

V. CONCLUSION

Each classifier has its own significance in text classification, but preprocessing of data is equally imperative which is base of more accuracy. In the proposed work we have taken specific terms in spite of all terms. In this paper performance of three classifiers are measured and finally compared them on the basis of many metrics like accuracy, f-measure, recall, precision and execution time.SVM classifier performed well according to the input datasets and give good results. The execution time of SVM is more than k-NN approach but this can be ignored because accuracy in classification is more preferred as compared to execution time.

REFERENCES

1. M.Ruiz, W.Lam and P. Srinivasan, “Automatic text categorization and its application to text retrivel,” IEEE Transaction Knowl. Data Engineering, Volume-11, November-Dec,1999, pp-865-879.
2. F.Sebastiani, “Machine Learning in automated text categorization,” ACM Computational Surveys, volume-34, 2002, pp. 1-47.
3. G.Forman, “An Extensive empirical study of feature selection metrics for text classification,” The Journ. Of Mc. Learn.(ML) Res., 2003, vol-3, pp-1289-1305.
4. Deepanshu, Dr. Ramesh Kait, “Document Classification using Svm combined with optimal feature selection,” Int. Journ. of Computer Engg. & Tech.(IJCET), may-june-2018, volume-9, issue-3, p.pages-250-258.
5. Semih E.,Serkan Gunal., M.Bilginer G., and O. Nezih G., “On feature extraction for spam e-mail detection”, In MM. Content Rp., Classification and security, Springer, 2006,pp-635-642.
6. Jianhua Guo, Guozhong Feng, Bing-Yi Jing and Lizhu hao, “A Bayesian feature selection paradigm for text classification,” Info. Processing & Mgt., 2012, 48,2,pp-283-30.
7. Yuefen Wang, Songbo T. and Gaowei W.,” Adapting centroid classifier for document categorization”, Expert system with app. 38,8, 2011, 10264-10273.
8. Alper Kursat Uysal and Serkan Gunal, “The impact of preprocessing on text classification”, Info. Processing & mgt. 50,1, pp-104-112, 2014.
9. JJ. Webster and Chunyu Kit, “Tokenization as the initial phase in NLP”, In Proc. of the 14th conf. on Computational Ling., Volume-4, Association for computational linguistics, 1992, p.p-1106-1110.
10. Miriam F., Hassan Saif, Yulan H. and Harith A., “On stopwords, filtering and data sparsity for sentiment analysis of twitter,” 2014.
11. David A hull et al., “Stemming algorithms: A Case study for detailded evaluation”, JASIS, 1996, 47,1, pp.-70-84, 2005.
12. Andreas N., Andreas Hotho and Gerhard PaaB, “A Brief Survey of Text Mining”, In Ldv Forum, ,2005, volume.-20 pp.-19-62.
13. Mike James, “Classification algorithm,” Wiley-Interscience, 1985.
14. Deepanshu, Ramesh Kait, “A Technical Review: Text Classification and Related approaches”, in proceedings of Int. Conf. on S& Tech.: Trends and Challenges(ICSTTC-2018).

15. A.Basu and M.Shephard, C.Waters, “Support Vector Machines for Text Categorization”, Proc. of the 36th Annual Hawaii Int. Conf. on Sys. Sc., 2003.
16. T.Joachins, “Text categorization with support vector machines: Learning with many relevant features”, in Proc. 10th European Conf. Mach. Learning(ECML),1998, pp.-137-142.
17. D.Lewis, “Naïve Bayes at Forty: The Independence Assumption in Information Retrivel”, Proc. of the 10th European Conf. on Mc. L.(ECML-98), 1998.
18. A.K. jain and Y.H.Li, “Classification of text documents,” The Compt. Journal,1998, volume-41, no.-8, p.pages-537-546.
19. Bo.Tang and H.He, “ENN:Extended nearest neighbor method for pattern recognition,” IEEE Computational, Intell. Mag.,2015, volume-10, no.3, pp. 52-60, 2015.
20. Bo T., Haibo H., Paul M. Baggenstoss and Steven Kay,” A Bayesian Classification Approach Using Class-Specific Features for Text Categorization” IEEE Trans. On Knowl. And Data Engg., Vol.28, No.-6, June 2016.
21. Dino Isa, Lam H. l., V.P Kallimani, and R.Raj K., “Text Documents Processing with the Bayes Formula for Classification using the Support vector machine,” IEEE Trans. Of Know. And Data Engg., 2008, vol.20, no.9,pp.1264-1272.
22. http: //en.wikipedia.org/w/ index.php? title=Text mining old id= 778865797
23. K.Rajeswari, VV., Kapil T., R.P. “Comparison of Different Classification Techniques using Different Datasets, Int.J. of Advances in Engg.& Tech., 2013, ISSN: 22231-1963.

AUTHORS PROFILE

Deepanshu, She has done B.tech (Computer Engineering) and M.tech (computer science and applications) with good academic score from Kurukshetra University Kurukshetra. She has published her paper in many well-reputed journals. Her recent publication is in IJCET, 2018.& her research work is in Data Science. She has good knowledge of Android application development and done many projects in java language.