INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume3, Issue3)
Available online at www.ijariit.com

Novel Approach to Text Classification by SVM-RBF Kernel and Linear SVC

Gurvir Kaur
Shaheed Udham Singh Engineering College, Tangori (Mohali)
Gurvir.kaur1990@gmail.com

Er. Parvinder Kaur
Shaheed Udham Singh Engineering College, Tangori (Mohali)
parvinderkaurcse@sus.edu.in

Abstract: In suspicion, characteristic dialect handling is an exceptionally brilliant strategy for the human-PC interface. Regular dialect grateful is once in a while alludes to as an Artificial Intelligence-whole issue since normal dialect distinguishing proof appears to include wide learning about the outside world and the capacity to control it. NLP has vital have regular qualities with the field of computational semantics and is frequently viewed as a sub-field of computerized reasoning. In this paper working two learning approaches knn and support vector machine (SVM) yet SVM gives importance great exactness, accuracy, review than KNN, SVC.

Keywords: NLP, KNN, SVC, TEXT Classification.

I. INTRODUCTION

Natural Language Processing (NLP) is defined as a field of computer science and linguistics which is concerned with the interactions between computers and human (natural) languages. In assumption, natural-language processing is a very smart method of the human-computer interface. Natural-language appreciative is sometimes refers to as an Artificial Intelligence-entire problem because natural-language identification seems to involve broad knowledge about the outside world and the ability to manipulate it. NLP has important have common characteristics with the field of computational linguistics and is often considered a sub-field of artificial intelligence.

Text mining is also recognized as data mining, refers to the procedure of deriving high worth information from text. The principle of data mining is a progression of raw and unstructured information; take out meaningful information from text. It generally involves the method of structuring the contribution text, deriving patterns contained by the structured data, and in conclusion evaluation and analysis of the output.

In text mining process, initializing with the gathering of documents, a tool extract the particular information or document and preprocess it. Then it comes to the next phase i.e. text analysis phase, where sometimes technique used is repeated until the relevant information is extracted. Text mining and data mining are similar except the data mining tools which organize the structured data from the databases only whereas text mining extract information from semi-structured and structured datasets such as HTML files, e-mails etc. therefore text mining is better option to handle or organize online data for companies.

Text Clustering (or Document Clustering) is defined as a procedure of cluster analysis to textual documents. It has a function in automatic document association, topic extraction, and fast information recovery. Text clustering is typically measured to be a centralized method. Text clustering consists of web document clustering for search users is an example of text clustering. The text clustering can be of two types, online and offline. Online are generally controlled by efficiency problems when compared to offline applications.

The foundation of text categorization goes back to early 60's, but it became a major subfield in the early 90's. In the late 90's machine learning approaches were effectively useful for categorization of text. In 1998 SVM technique was used for categorization of text. In 1999 Maximum Entropy models were applied. In 1999 EM algorithm was also proposed by McCallum to train a hybrid model. In 2000 Multi-label classification is investigated of multiple topics and AdaBoost was used to enhance the multi-labels classification. In 2005 Maximum Entropy models were extended to a new version multi-labeled Max Entropy Models for text classification. Content-based document management tasks had gained a major role in the information system field in last ten years (2006-2015), due to the increased availability of documents.

With the fast development of online information efficient recovery of several exacting information is complicated without excellent indexing and summarization of document content. To handle and organize the huge text collection, Text Classification may be the solution in text mining. Text Classification, also known as text categorization, is defined as the task of defining unlabeled documents into predefined classes automatically.

II. LITERATURE REVIEW

Wen Zhanget.al [1]: - In this paper author describes two tasks of text representation i.e. term weighting and indexing. In this paper author also have done a comparison of three methods (TF-IDF, LSI, multi-word) for text representations. In this paper, two documents i.e. Chinese and English are used to evaluate the text representations method for text classification. The main objective of this paper is to study the efficiency and effectiveness of different text representation methods. For text representation, two main tasks are indexing which is mainly concerned with statistical and semantic quality and weighting which is mainly concerned with term frequency (TF) and inverse document frequency (IDF).

VishwanathBijalwanet.al [2]: - In this paper, KNN based learning approach is used for text categorization. Text categorization is the task of automatically allocating unlabeled documents into predefined categories. If a document or text belongs to exactly one class or category, it is known as single-label classification task and if a document or text belongs to more than one or more class or category, it is known as a multi-label classification task. The author firstly classifies the documents using KNN based machine learning approach and then compared with Naïve Bayes and Term-graph approach by returning the most relevant documents. In this paper, the author concludes that KNN shows the maximum accuracy as compared to the Naive Bayes and Term-Graph.

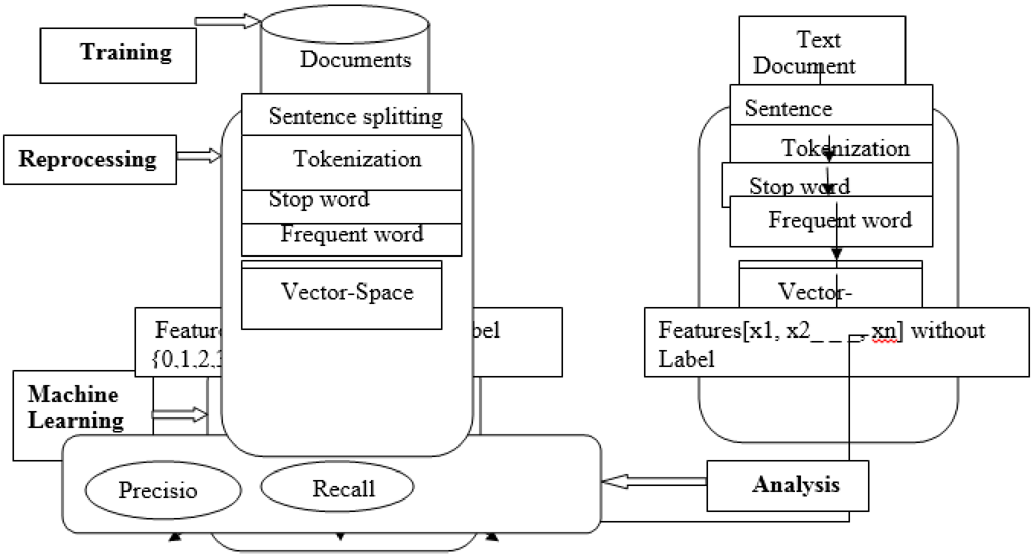
Aixin Sunet.al [3]: -In this paper, the author purposed a simple, scalable and non-parametric approach for short text classification. In general short texts are much nosier, shorter and sparser therefore to improve short text representation author proposed to trim a short text categorization. This approach mimics human classification process for a piece of short text like tweets, status updates, and comments. It selects the representative words from a given short text as query words. After that, it searches for a set of labeled text those best matches the query words. The author has used four approaches and is evaluated to select the query words: TF, TF.IDF, TF.CLARITY and TF.IDF.CLARITY. The proposed approach achieves accuracy with the baseline Maximum Entropy classifier. Experimental results show that TF.CLARITY performs effectively when three or more words are used in a query whereas TF.IDF.CLARITY performs well when one word is used in a query. The improvement becomes very minor when more than five words are used in a query.

Mengen Chenet.al [4]: - In this paper, short text classification is optimized by learning multi-granularity topics. Author has proposed a new algorithm using multi-granularity to create features for short text. Due to sparsity and shortness of short text, it is different from usual documents. Two major approaches used for short text classification to improve the representation of short text are:-
i) Fetch the appropriate information of short text to directly add more text.
ii) Drive latent topics from existing large corpus.

TanmayBasuet.al [5]: - Text classification is a difficult task due to its high dimensionality of data. Therefore, efficient method for feature selection is required to improve the performance of text classification. This paper presents a new feature selection method for text classification using supervised term selection approach. In this paper TS (term significance) a feature selection technique is compared with CHI, IG & MI. The proposed approach derives a similarity score between a term and a class and then ranks the terms according to their scores over all the classes. The experimental results show that the proposed TS can produce better classification accuracy even after removing 90% unique terms.

METHODOLOGY

1. First preprocess the documents and text documents by split the sentences, tokenization, remove stop words and to find the frequent words.
2. Find the weight by using TF-IDF (Term frequency- inverse document frequency).
3. Next, make the Vector Space Model
4. After that features can be labeled with classes on the other hand in the text document features cannot be labeled with classes.
5. In Machine, learning part apply SVM-RBF (Support Vector Model-Radial Basis Function) algorithm on each class.
6. SVM-RBF Model can be implemented by using the algorithm.
7. In the Analysis, phase verifies that the result which can be given by SVM-RBF Model can be accurate, precision or recall.



III.ssRESULTS AND DISCUSSIONS

Table and Graphs

Analysis the text classification on the basis of precision, recall, and accuracy, In the analysis we make different table and graph for precision, recall and accuracy for K-mean clustering and SVM RBF Kernel

Comparison

The conclusion of above experiment SVM-RBF Kernel better texts classification on the basis of precision, recall, and accuracy.

Table no.4: Result Table for Comparison between SVC and SVM-RBF Kernel

Training Instance	Precision (Linear svc)	Precision (SVM-RBF Kernel)	Recall (Linear SVC)	Recall (SVM-RBF Kernel)	Accuracy (Linear SVC)	Accuracy (SVM-RBF Kernel)
60	87.75	93	88.5	94.25	93.25	96.25
65	88.25	93.5	89	94.5	94	96.5
70	87.5	95.25	88.5	96	93.5	97.25
75	86.5	96	88	96.5	93.25	97.87
80	87.25	97.5	89	97.5	93.25	98.5
85	85.5	97.5	87.75	97.5	92.75	98.5
90	86.25	98.75	87.5	98.75	92.75	99

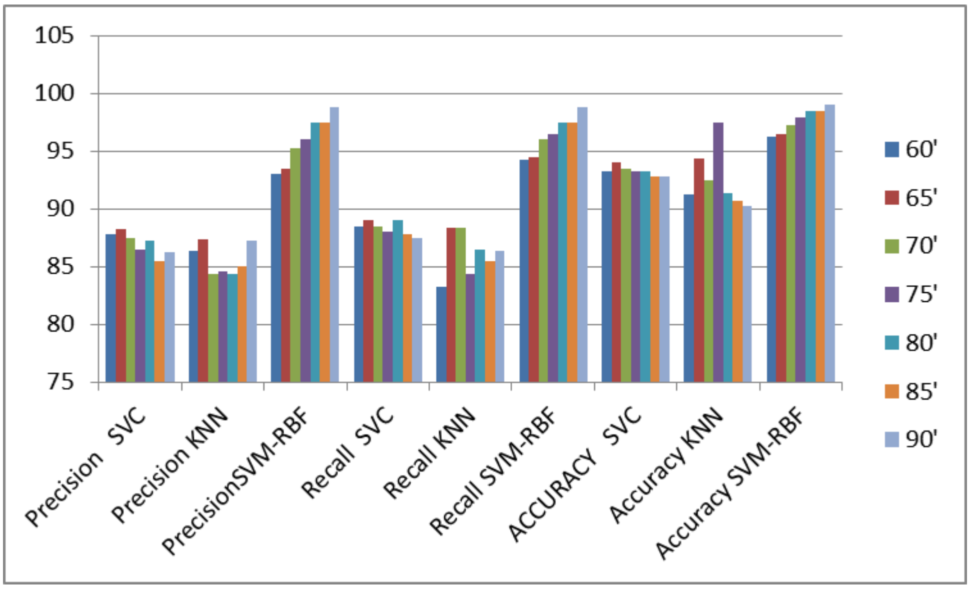


Figure: Comparison Result Graph between SVC, KNN and SVM-RBF Kernel

CONCLUSIONS

Above tables show the best result for best two algorithms k-mean and SVM. Considering the low false positive ratio knn performs well as it is easier to implement and has low running time but has less accuracy than Linear SVM. Hence we conclude that optimization methods perform well and show better results than other classifiers. We enhance this work by using a Graphical model like a conditional random field, hidden markov field which shows the dependency between the features in text classification. We can also use Kernel function for reducing the processing time and error of overlapping information in text classification. The discriminative approaches make assumptions of their own that are subject to violations.

REFERENCES

[1] Wen Zhang, Taketoshi Yoshida, Xijin Tang “A comparative study of TF*IDF, LSI and multi-words for text classification”, *Elsevier in advances engineering software*, Vol.38,
Pp.2758-2765, 2011.

[2] VishwanathBijalwan, Vinay Kumar, PinkiKumari, Jordan Pascual“ KNN based Machine Learning Approach for Text and Document Mining”, *International Journal of database theory and application*, Vol.7,No.1,pp.61-70, 2014.

[3]Aixin Sun, “Short Classification using very few words”, *International Journal of computer applications technology and research*, Vol.2, pp.4503-1475, 2012.

[4]Mengen Chen, Xiaoming Jin, Dou Shen, “Short Text Classification Improved by Learning Multi-Granularity Topics”, an I*nternational joint conference on artificial intelligence*, Vol.1, pp.1776-1781, 2010.

[5] TanmayBasu, C. A. Murthy, “Effective Text Classification by a Supervised Feature Selection Approach”, *International conference on machine learning*, Vol.14, pp.1289-1297, 2008.

[6] YoungjoongKo, “A Study of Term Weighting Schemes Using Class Information for Text Classification”, *International Journal of computer applications technology and* research, Vol.2, pp.4503-1475, 2012.

[7]Svetlana Kiritchenko, Stan Matwin, “Email Classification with Co-training”, *International Journal of computer applications,* Vol.56, No.10, pp.1-10, 2006.

[8] Guansong Pang, Shengyi Jiang, “ A Generalized Cluster Centroid-based classifier for text categorization”*Expert Systems with Applications*, Vol.8,pp.2758-2765,2013.

[9]M. Ikonomakis, S. Kotsiantis, V. Tampakas, “Text Classification Using Machine Learning Techniques”, *Wseas transactions on computers*, Issue 8, Volume 4, pp. 966-974, 2005,

[10] Sang Min Lee, Dong SeongKim, Ji Ho Kim,JongSou Park, “Spam Detection Using Feature Selection and Parameters Optimization”, *IEEE*, pp. 883-888, 2010.

[11] SarwatNizamani, NasrullahMemon, UffeKockWiil, Panagiotis Karampelas, “Modeling Suspicious Email Detection using Enhanced Feature Selection”, *Wall Street Journal*, Vol.14, April 2012.

[12] QianXu, Evan Wei Xiang and Qiang Yang, “SMS Spam Detection Using Non-Content Features”, *IEEE Intelligent systems*, Vol.27, No. 6, pp. 44-51, Nov.-Dec. 2012.

[14] M. Mangalindan, “For bulk E-mailer, pestering millions offers a path to profit,” *Wall Street Journal*, Vol.13, pp.7-13, November 13, 2002.

[15] IsmailaIdris, AbdulhamidShafi’i Muhammad, “An Improved AIS Based E-mail Classification Technique for Spam Detection”, *The Eighth International Conference on eLearning for Knowledge-Based Society*, Vol. 6318, pp.20-24, February 2012, Thailand