IOSR Journal of Engineering (IOSR JEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 06-09

Text Categorization using Clustering and Classification Machine Learning Algorithms via NLP

PatilKiran Sanjay[1] Prof.Kurhade N.V[2]
P.G. Student, Department of Comp Engineering SharadchandraPawar college of Engineering, Otur, Pune
Professor, Department of Comp Engineering, SharadchandraPawar college of Engineering, Otur, Pune

Abstract: *In a world that routinely delivers progressively textual information. It is basic errand to dealing with that printed information. There are numerous content investigation strategies are accessible to overseeing and imagining that information, however numerous procedures may give less precision on account of the uncertainty of natural language. To give the fine-grained investigation, in this paper present efficient machine learning algorithms for classify content information. To enhance the precision, in proposed framework I acquainted NLTK python library with perform natural language processing. The principle point of proposed framework is to sum up the model for ongoing application by utilizing efficient text classification as well as clustering algorithms and find the precision of model utilizing execution measure.*
Keywords:*Text analytics, TF-IDF, Text classification, Text categorization.*

I. Introduction

With the rapid growth of on line information, text categorization has become one of the key techniques for handling and organizing text data. Text categorization techniques are used to classify news stories, to find interesting information on the WWW, and to guide a user's search through hypertext. Since building text classifiers by hand is difficult and time-consuming. In this paper I will explore and identify the benefits of different type of techniques like classification and clustering for text categorization. Here I have labeled as well as non labeled data for analysis by using supervised as well as unsupervised machine learning algorithms I can categorized the data efficiently and after text categorization I will compare all techniques and visualized which is better for real time applications.

The main purpose of proposed system is that create generalized model as per user's requirements, because when we apply machine learning algorithms on dataset then they gives different result.

Before going to categorize the dataset we have to apply preprocessing on that data and then pass that data preprocessing output to classification or clustering algorithms as a input. For data preprocessing here I have used natural language processing (NLP).

II. Literature Survey

A **According to Divyansh Khanna, Rohan Sahu, Veeky Baths, and Bharat Deshpande[2]** This study provides a benchmark to the present research in the field of heart disease prediction. The dataset used is the Cleveland Heart Disease Dataset, which is to an extent curated, but is a valid standard for research. This paper has provided details on the comparison of classifiers for the detection of heart disease. We have implemented logistic regression, support vector machines and neural networks for classification. The results suggest SVM methodologies as a very good technique for accurate prediction of heart disease, especially considering classification accuracy as a performance measure. Generalized Regression Neural Network gives remarkable results, considering its novelty and unorthodox approach as compared to classical models.

From this I had taken the idea of SVM algorithm for classification.

According to KrunoslavZubrinic, Mario Milicevic and IvonaZakarija[3] In this research we tested the ability of classification of CMs using simple classifiers and bag of words approach that is commonly used in document classification. In two experiments we compared the results of classification randomly selected CMs using three classifiers. The best results are achieved using multinomial NB classifier. On reduced set of attributes and instances that classifier correctly classified 79.44 of instances. We believe that the results are promising, and that with further data preprocessing and adjustment of the classifiers they can be improved.

From this this I had introduced NB classifiers algorithm in my system for mapping the different datasets.

According to Thorsten Joachims This [4]paper introduces support vector machines for text categorization. It provides both theoretical and empirical evidence that SVMs are very well suited for text categorization. The theoretical analysis concludes that SVMs acknowledge the particular properties of text:
1. high dimensional feature spaces
2. few irrelevant features (dense concept vector)
3. sparse instance vectors.

The experimental results show that SVMs consistently achieve good performance on text categorization tasks, outperforming existing methods substantially and significantly. With their ability to generalize well in high dimensional feature spaces, SVMs eliminate the need for feature selection, making the application of text categorization considerably easier. Another advantage of SVMs over the conventional methods is their robustness. SVMs show good performance in all experiments, avoiding catastrophic failure, as observed with the conventional methods on some tasks. Furthermore, SVMs do not require any parameter tuning, since they can find good parameter settings automatically. All this makes SVMs a very promising and easy-to-use method for learning text classifiers from examples.

According to Payal R. Undhad,Dharmesh J. Bhalodiya[5] Text classification is a data mining technique used to predict categorical label. Aim of research on text classification is to improve the quality of text representation and develop high quality classifiers. Text classification process includes following steps i.e. collection of data documents, data preprocessing, Indexing, term weighing methods, classification algorithms and performance measure. Machine learning techniques have been actively explored for text classification. Machine learning algorithm for text classification are Naive Bayes classifier, K-nearest neighbor classifiers, support vector machine. Text classification is very helpful in the field of text mining, The volume of electronic information is increase Day by Day and its extracting knowledge from these large volumes of data. The classification problem is the most essential problems in the machine learning along with data mining literature. This paper survey on text classification. This survey focused on the existing literature and explored the documents representation and an analysis classification algorithms Term weighting is one of the most vital parts for construct a text classifier. The existing classification methods are compared based on pros and cons. From the above discussion it is understood that no single representation scheme and classifier can be mentioned as a general model for any application Different algorithms perform differently depending on data collection.TF-IDF word embedding concept is taken from this paper for vectorization.

According to Deokgun Park, Seungyeon Kim, Jurim Lee, JaegulChoo, Nicholas Diakopoulos, and NiklasElmqvist[1] Current text analytics methods are either based on manually crafted human-generated dictionaries or require the user to interpret a complex, confusing, and sometimes nonsensical topic model generated by the computer. In this paper we proposed Concept Vector, a novel text analytics system that takes an visual analytics approach to document analysis by allowing the user to iteratively defined concepts with the aid of automatic recommendations provided using word embedding. The resulting concepts can be used for concept-based document analysis, where each document is scored depending on how many words related to these concepts it contains. We crystallized the generalizable lessons as design guidelines about how visual analytics can help concept based document analysis. We compared our interface for generating lexica with existing databases and found that Concept Vector enabled users to generate concepts more effectively using the new system than when using existing databases. We proposed an advanced model for concept generation that can incorporate irrelevant words input and negative words input for bipolar concepts. We also evaluated our model by comparing its performance with a crowd sourced dictionary for validity. Finally, we compared Concept Vector to Empath in an expert review. The text analysis provided by Concept Vector enables several novel concept-based document analysis, such as richer sentiment analysis than previous approaches, and such capabilities can be useful for data journalism or social media analysis. There are many limitations that Concept Vector does not solve. Among these, the selection/integration of multiple heterogeneous training data according to the target corpus and the automatic disambiguation of multiple meanings of words according to the context are promising avenues of future research.
In proposed system I introduced text categorization on labeled and non labeled data to create generalized model for real time applications.

III. Problem Statement

The proposed work is on textual dataset, using classification and clustering machine learning algorithms perform text categorization. If data is labeled then text categorization is using classification otherwise using clustering ML algorithm and find the best algorithm for input dataset by using performance measure.
The main purpose of this system is to provide generalized model for real time applications.

Objectives of System

- To provides generalized model for real time applications.
- To categorized large labeled as well as non labeled textual dataset efficiently.
- To applying different ML algorithm for different dataset and find accuracy of model using performance measure.

Scope of System

- To provides efficient text categorization.
- To provide great user experience to users in their day to day activity this text categorization to be analyzed.

IV. Proposed System

In today's world, most of work is doing on textual data. Huge textual data is very critical to handle, for maintaining that textual data here used some machine learning algorithms. If data is labeled then it can handle using classification ML algorithms like SVM, Naive Bayes.

If data is not labeled then this type of textual data is group by using clustering ML algorithms like K-means, Guassian Mixture Model.

After applying algorithms the main aim of proposed system is to find the efficient ML algorithm for particular input dataset using performance measure.

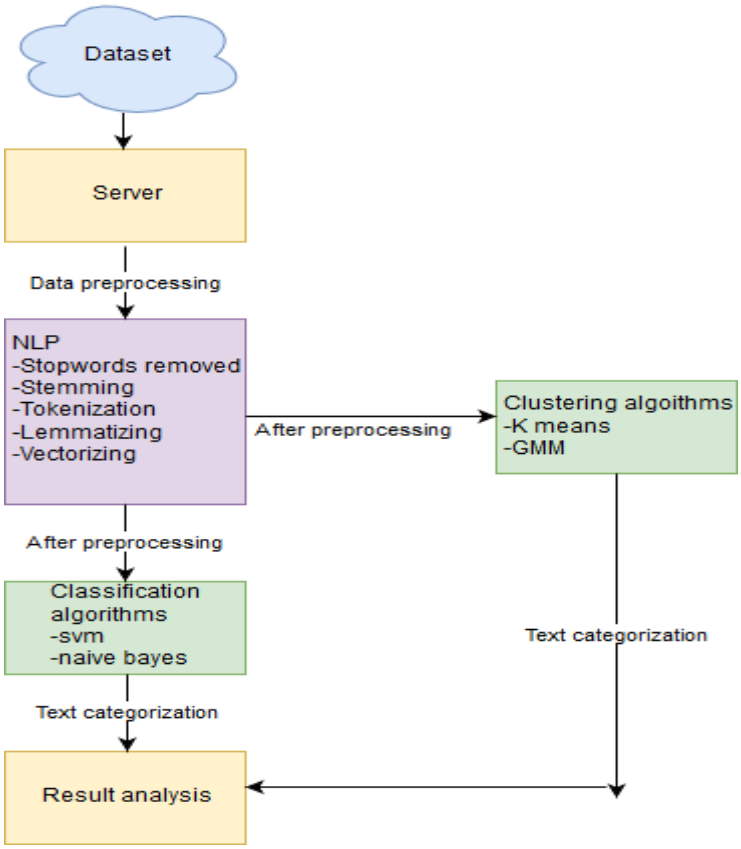

Figure 1:Proposed System Architecture

V. Conclusion

In this research work, the main focus is on the text categorization, whenever data is labeled or unlabeled by using machine learning algorithms classify free text efficiently. Support vector machine (SVM) and naive Bayes classification algorithm for labeled data and K-means and Gaussian mixture model (GMM) clustering algorithm for non-labeled data.

The main purpose of this project is to map any real time text categorized problem to appropriate machine learning algorithm and find accurate confidence probability of data item. Efficiency of machine learning algorithm is varying with each dataset. By using performance measure calculate the accuracy model for classification. After that I will visualized that result using python libraries.

Future Work

Using MD5 algorithm we can calculate 100% accuracy of SVM algorithm.

References

[1]. Deokgun Park, et al. "Concept Vector: Text Visual Analytics via Interactive Lexicon Building using Word Embedding", IEEE Transactions on Visualization and Computer Graphics, Vol. 24, NO. 1,2018
[2]. Divyansh Khanna, et al. "Comparative Study of Classification Techniques (SVM, Logistic Regression and Neural Networks) to Predict the Prevalence of Heart Disease" International Journal of Machine Learning and Computing, Vol. 5, No. 5, October 2015.
[3]. KrunoslavZubrinic,et al "Comparison of Naive Bayes and SVM Classifiers in Categorization of Concept Maps" International Journal of computers Issue 3, Volume 7, 2013
[4]. Thorsten Joachims "Text Categorization with Support Vector Machines :Learning with Many Relevant Features"
[5]. Payal R. Undhad, Dharmesh J. Bhalodiya , "Text Classification and Classifiers: A Comparative Study" 2017 IJEDR,Volume 5, Issue 2,ISSN: 2321-9939
[6]. M. Berger, K. McDonough, and L.M. Seversky. "cite2vec: Citation driven document exploration via word embeddings." IEEE Transactions on Visualization and Computer Graphics, 23(1):691700, Jan 2017.
[7]. https://www.nltk.org/book/
[8]. Lkit:A Toolkit for Natuaral Language Interface Construction