International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume: 3 | Issue: 4 | May-Jun 2019 Available Online: www.ijtsrd.com e-ISSN: 2456 - 6470

Experimental Result Analysis of Text Categorization using Clustering and Classification Algorithms

Patil Kiran Sanajy, Prof. Kurhade N. V.

Department of Comp Engineering, Sharadchandra Pawar College of Engineering, Otur, Pune, India

How to cite this paper: Patil Kiran Sanajy | Prof. Kurhade N. V. "Experimental Result Analysis of Text Categorization using Clustering and Classification Algorithms" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-3 | Issue-4, June 2019, pp.1216-1219, URL: https://www.ijtsrd.com/papers/ijtsrd25077.pdf

IJTSRD25077

ABSTRACT

In a world that routinely produces more textual data. It is very critical task to managing that textual data. There are many text analysis methods are available to managing and visualizing that data, but many techniques may give less accuracy because of the ambiguity of natural language. To provide the fine-grained analysis, in this paper introduce efficient machine learning algorithms for categorize text data. To improve the accuracy, in proposed system I introduced Natural language toolkit (NLTK) python library to perform natural language processing. The main aim of proposed system is to generalize the model for real time text categorization applications by using efficient text classification as well as clustering machine learning algorithms and find the efficient and accurate model for input dataset using performance measure concept.

Keywords: *Text analytics, Term frequency–Inverse document frequency (TF-IDF), Text classification, Text categorization*

INTRODUCTION

Now a day's most probable work is on huge amount of text data, text categorization has become one of the important methods for handling and organizing text data. Text categorization techniques are used to classify news stories, to find interesting information on the internet, and to guide a user's search through hypertext. Since building text classier by hand is troublesome and tedious.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 input. For data preprocessing here I have used natural language processing (NLP).

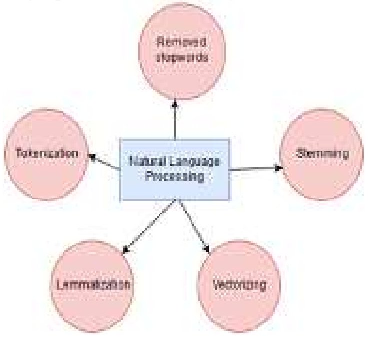

Figure 1: Natural Language Processing

Removing stop words: Stop words are regular words that show up in each archive they have small importance, they serve just syntactic significance yet don't demonstrate subject make a difference it is all around perceived among the compliance recovery specialists that a lot of practical English words (eg. the, an, and, that, this, is, an) is pointless as ordering terms. These words have low Discrimination esteem, since they happen in each English report. Henceforth they don't help in recognizing archives about different subjects. The way toward evacuating the arrangement of bearing utilitarian words from the arrangement of words created by word extraction is known as stop words expulsion. So as to expel the stop words, first step is making a rundown of stop words to be evacuated, which is additionally called as the stop word list. After this, second step is the arrangement of words created by word extraction is then examined with the goal that each word showing up in the stop list is evacuated.

Stemming: In stemming different types of a similar word are changed over into a solitary word. For instance, particular, plural, and different tenses are changed over into a solitary word. Port stemmer calculation is notable calculation for stemming. e.g. connection to connect, computing to compute.

Tokenization: Tokenizing separates text into units such as sentences or words. It gives structure to previously unstructured text. e.g. Plata o Plomo – 'Plata', 'o', 'Plomo'.

Lemmatizing: Lemmatizing derives the canonical form (lemma) of a word. i.e the root form. It is better than stemming as it uses a dictionary based approach i.e a morphological analysis to the root word. e.g. Entitling, Entitled-Entitle

LITARATURE SURVEY

A **According to Divyansh Khanna, Rohan Sahu, Veeky Baths, and Bharat Deshpande[2]** This examination gives a benchmark to the present research in the field of heart disease prediction. The dataset utilized is the Cleveland Heart Disease Dataset, which is to a degree curated, yet is a substantial standard for research. This paper has given subtleties on the correlation of classifiers for the discovery of heart disease. We have executed strategic relapse, bolster vector machines and neural systems for arrangement. The outcomes propose support vector machine (SVM) philosophies as a decent strategy for exact prediction of heart disease, particularly considering grouping exactness as an execution measure. Summed up Regression Neural Network gives momentous outcomes, thinking about itscuriosity and unconventional methodology when contrasted with established models.

From this I had taken the idea of support vector machine (SVM) algorithm for classification.

According to Krunoslav Zubrinic, Mario Milicevic and Ivona Zakarija[3] In this research we tested the ability of classification of concept map (CM)s 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 Naive Bayes 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 Naive Bayes classifiers algorithm in my system for mapping the different datasets.

According to Thorsten Joachims This [4] paper presents support vector machines for text categorization. It gives both hypothetical and exact proof that support vector machine (SVMs) are very appropriate for text categorization. The hypothetical investigation reasons that SVMs recognize the specific properties of text:

1. high dimensional feature spaces
2. few irrelevant features
3. sparse instance vectors.

The experimental results demonstrate that SVMs reliably accomplish great execution on text categorization undertakings, beating existing techniques considerably and altogether. With their capacity to sum up well in high dimensional element spaces, SVMs dispose of the requirement for highlight determination, making the utilization of text categorization impressively less demanding. Another favorable position of SVMs over the ordinary strategies is their vigor. SVMs show great execution in all trials, dodging disastrous disappointment, as saw with the ordinary techniques on a few errands. Besides, SVMs don't require any parameter tuning, since they can find great parameter settings consequently. This makes SVMs a promising and simple to-utilize strategy for taking in text classifiers from precedents.

According to Payal R. Undhad,Dharmesh J. Bhalodiya[5] Text classification is an information mining procedure used to foresee clear cut name. Point of research on text classification is to enhance the nature of text portrayal and grow superb classifiers. Text classification process incorporates following advances for example accumulation of information records, information preprocessing, Indexing, term gauging strategies, classification calculations and execution measure. Machine learning strategies have been effectively investigated for text classification. Machine learning calculation for text classification are Naive Bayes classifier, K-closest neighbor classifiers, bolster vector machine. Text classification is useful in the field of text mining, The volume of electronic data is increment step by step and its extricating information from these huge volumes of information. The classification issue is the most basic issues in the machine learning alongside information mining writing. This paper overview on text classification. This review concentrated on the current writing and investigated the reports portrayal and an examination classification calculations Term weighting is a standout amongst the most imperative parts for build a text classifier. The current classification strategies are analyzed dependent on advantages and disadvantages. From the above discourse it is comprehended that no single portrayal plan and classifier can be referenced as a general model for any application Different calculations perform contrastingly relying upon information gathering.

Term frequency–Inverse document frequency (TF-IDF) word embedding concept is taken from this paper for vectorization.

According to Deokgun Park, Seungyeon Kim, Jurim Lee, Jaegul Choo, Nicholas Diakopoulos, and Niklas Elmqvist[1] Current text analytics techniques are either founded on physically created human-produced word references or require the client to decipher a perplexing, confounding, and at times silly subject model produced by the computer. In this paper we proposed Concept Vector, a novel text analytics framework that adopts a visual analytics strategy to record examination by enabling the client to iteratively defined concepts with the guide of programmed proposals gave utilizing word inserting. The subsequent concepts can be utilized for concept-based archive investigation, where each record is scored relying upon what number of words identified with these concepts it contains. We solidified the generalizable exercises as plan rules about how visual analytics can help concept based record examination. We contrasted our interface for producing lexica and existing databases and found that Concept Vector empowered clients to create concepts more effectively utilizing the new framework than when utilizing existing databases. We proposed a propelled model for concept age that can consolidate unimportant words info and negative words contribution for bipolar concepts. We likewise assessed our model by contrasting its execution and a publicly supported word reference for legitimacy. At long last, we contrasted Concept Vector with Empath in a specialist audit. The text investigation given by Concept Vector empowers a few novel concept-based record examination, for example, more extravagant assessment investigation than past methodologies, and such capacities can be valuable for information reporting or internet based life investigation. There are numerous constraints that Concept Vector does not fathom. Among these, the determination / joining of numerous heterogeneous preparing information as indicated by the objective corpus and the programmed disambiguation of various implications of words as per the context are promising roads of future research. In proposed system I introduced text categorization on labeled and non-labeled data to create generalized model for real time applications.

OBJECTIVES OF SYSTEM

The Objective of the proposed application is as follows:

- 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.

PROPOSED METHODOLOGY

Text categorization by using supervised and unsupervisedmachine learning algorithms as follos:

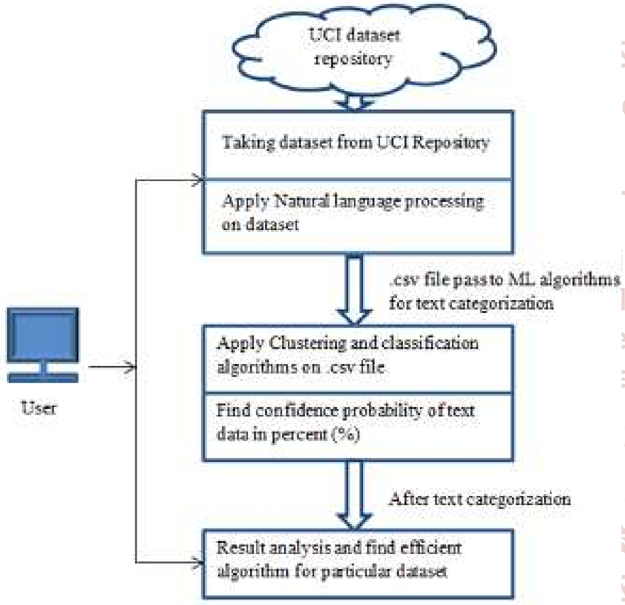
UCI dataset repository
Taking dataset from UCI Repository
Apply Natural language processing on dataset
.csv file pass to ML algorithms for text categorization
Apply Clustering and classification algorithms on .csv file
Find confidence probability of text data in percent (%)
After text categorization
Result analysis and find efficient algorithm for particular dataset
User

Figure 2: Proposed System Architecture

In ebb and flow investigate programmed classification [2] of reports into predefined classes has seen as a functioning consideration, the archives can be characterized in three different ways, unsupervised, supervised and semi supervised strategies. From most recent couple of years, the undertaking of programmed text classification has been broadly considered appears around there, including the machine learning methodologies, for example, Naive Bayes classifier, Support Vector Machines (SVMs).

Classification: When input (x1, x2…., xn) and output(y1,y2,….yn) is available and we have to mapped input set to output set using supervised ML algorithms.

- Support vector machine(SVM)
- Naive Bayes classification.

Clustering: When only input set is available (x1,x2…xn) then we have to group similar type of data depend on unsupervised machine learning algorithms. This text categorization technique is only for un- labeled data

- K-means clustering
- Guassian mixture model(GMM)

After applying machine learning algorithms then find out the appropriate technique for particular dataset by using performance measure.

SYSTEM ANALYSIS

Steps for Execution:

Input: Dataset D in the form of .csv file.
Output: Confidence probability of text data.
Step 1: Take dataset from UCI ML repository.
Step 2: convert into trained knowledge base dataset i,e csv file.
Step 3: csv file pass as a input to preprocessing module via NLP.
Step 4: pass output of preprocessing to machine learning algorithms as a input for performing text categorization.

- If data is labeled then used supervised Machine Learning means classification algorithms.
- If data is non-labeled then used unsupervised Machine Learning means clustering algorithms.

Step 5: after performing Machine Learning algorithms then find out confidence probability of text data.
Step 6: select appropriate algorithm for particular dataset depend on confidence probability.

RESULT AND DISCUSSION

In my research I have taken one dataset for both type of classification i.e. Tweet analysis dataset. when SVM(support vector machine) and naive bayes classification had apply on that dataset then naive bayes gives better result than SVM of text classification.

I have taken 10 records for comparison in SVM and naive bayes classification then result is shown below

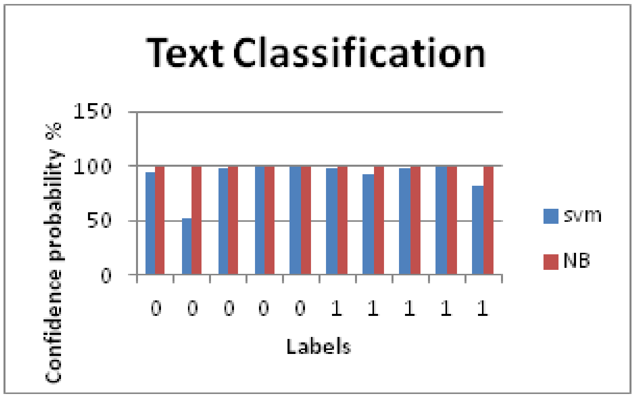

Figure 3: Comparison between SVM and NB for text

In fig. 3 X-axis shows the labels and Y-axis shows output i.e. confidence probability in percent(%)that means how many percent tweet text is to be good(1)orbad(0).Similarly, I have taken another dataset for both type of clustering i.e. Songs dataset, when Kmeans and Gaussian Mixture Model(GMM) clustering had apply on that dataset that time Kmeans gives centroid based result but if text data does not able to foundcentroid that time GMM works based on density of data. That's why GMM is better than Kmeans clustering because its applicable for all types of datasets.

On this above result I have conclude that machine learning algorithms are gives different result for different datasets. That's why we can apply ML algorithm on any dataset and find out which gives the better result.

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.

REFERENCES

[1] Deokgun Park, Seungyeon Kim, Jurim Lee and Jaegul Choo. "Concept Vector: Text Visual Analytics via Interactive Lexicon Building usingWord Embedding", IEEE Transactions on Visualization and Computer Graphics,Vol.24, IEEE, January 2018

[2] Divyansh Khanna, Rohan Sahu, Veeky Baths, and Bharat Deshpande. "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 2015, Vol.5,IJMLC, October 2015.

[3] Krunoslav Zubrinic, Mario Milicevic and Ivona Zakarija. "Comparison of Naive Bayes and SVM Classifiers in Categorization of Concept Maps" International Journal of computers, Vol.7 ,IEEE, 2013

[4] Thorsten Joachims. "Text Categorization with Support Vector Machines :Learning with Many Relevant Features"

[5] Payal R. Undhad and Dharmesh J. Bhalodiya , "Text Classification and Classifiers: A Comparative Study" International conference on IJEDR, Vol.5,2017

[6] M. Berger, K. McDonough, and L.M. Seversky. "cite2vec: Citation driven document exploration via word embeddings." IEEE Transactions on Visualization and Computer Graphics, January 2017.

[7] https://www.nltk.org/book/

[8] Lkit:A Toolkit for Natuaral Language Interface Construction