International Journal of Scientific Research in Science, Engineering and Technology (www.ijsrset.com)
© 2020 IJSRSET | Volume 7 | Issue 1 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
DOI : https://doi.org/10.32628/IJSRSET207117

Multi - Class Document Classification: Effective and Systematized Method to Categorize Documents

Kaushika Pal*[1], Dr. Biraj V. Patel[2]

*[1]Assistant Professor, Sarvajanik College of Engineering and Technology, Surat, Gujarat, India

[2]G. H. Patel, P.G. Department of Computer Science and Technology, Sardar Patel University, V.V. Nagar, Gujarat, India

ABSTRACT

A large section of World Wide Web is full of Documents, content; Data, Big data, unformatted data, formatted data, unstructured and unorganized data and we need information infrastructure, which is useful and easily accessible as an when required. This research work is combining approach of Natural Language Processing and Machine Learning for content-based classification of documents. Natural Language Processing is used which will divide the problem of understanding entire document at once into smaller chucks and give us only with useful tokens responsible for Feature Extraction, which is machine learning technique to create Feature Set which helps to train classifier to predict label for new document and place it at appropriate location. Machine Learning subset of Artificial Intelligence is enriched with sophisticated algorithms like Support Vector Machine, K – Nearest Neighbor, Naïve Bayes, which works well with many Indian Languages and Foreign Language content's for classification. This Model is successful in classifying documents with more than 70% of accuracy for major Indian Languages and more than 80% accuracy for English Language.

Keywords : Classification, NLP, Machine Learning, and Feature Set, Accuracy

I. INTRODUCTION

Data is widely spread on World Wide Web as we are generating so much data every day from mobile applications, Google Maps, uploading documents in various formats from various websites. Content in form of documents, posts, blogs on various issues, videos all are posted on web. In future this abandoned of data can be used to extract information for it's efficient usage and this is challenge for researchers to extract information using Artificial Intelligence. The data on which this research work is focusing is documents. The very first step for deriving information is to have related data classified and clustered at one place. This research work is discussing techniques and methodology to classify documents and label all on the basis on content in the documents, using Natural Language Processing and Machine Learning techniques, which is subset of Artificial Intelligence. Classification is also possible on file names, but this research work will process and classify documents based on its content.

The way we humans learn by experiences, and can actually predict or visualize what can be the outcome based on our experience; similarly machine can be trained by providing them with experienced data to learn and then provide them with new data to predict the outcome.

This work is using Machine Learning, which learn from examples. Figure 1 shows an overview of relation between Artificial Intelligence and Machine Learning.

Artificial Intelligence deals for simulation of intelligent behavior in machines, which is capable of, replicate human behavior and the machine is computer.

Machine Learning also said to be an application of Artificial Intelligence provides systems the ability to learn from experiences without being programmed. It's like give them experiences and from past experiences they will learn and ready to predict the outcome of new situation.

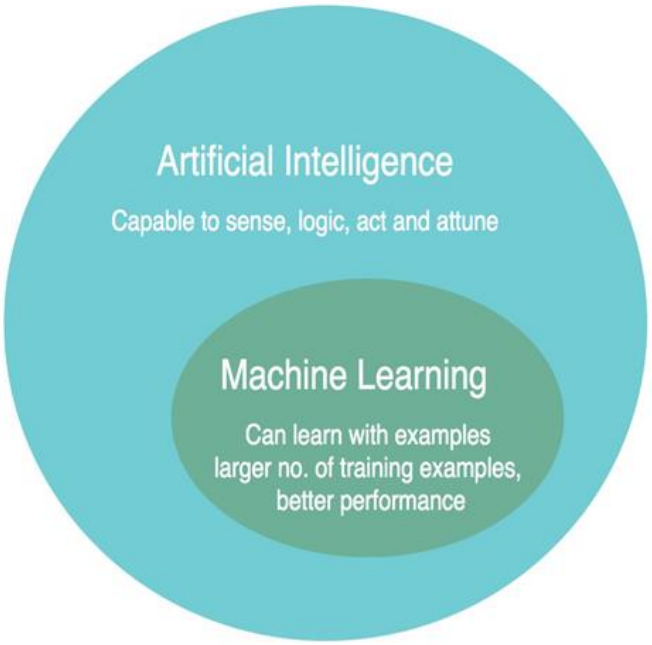

Figure 1. Machine Learning subset of Artificial Intelligence.

II. LITERATURE SURVEY

Jayashri K, et al., have used machine learning to classify sentiments into positive and negative for opinion mining. [1] The authors represented use of Naïve Bayes and Support Vector Machine for classification of 2 classes. J. Kaur, et al. tested 10 Machine Learning algorithms for classifying Sentiments in Punjabi Language [2] authors used algorithms to find the best one, which suits Punjabi text and found SVM suits better for Punjabi data. Harikrishna D M et al., classified children stories based on story structure [3] they used SVM, KNN and NB algorithms. K. Pal et al, proposed a model to classify Hindi poem into navrasas [4] the authors are using emotional and poetic feature to categories poems into 9 classes. Shalini P. et al, have proposed a classifier to classify textual imaged documents into categories [5] the authors proposed a survey for the feasibility of the system using statistical measurements of Hindi keywords obtained from different sources and by finding the challenges which will be faced while classifying printed and handwritten documents. Shalini P. et al, proposed a model to classify Hindi documents into predefined classes [6] they notice the challenges of the language and tested model with only 4 documents of 2 categories with 100% accuracy. K. Pal et al, have carried out a survey on classification task of Indic languages [7] they found that using different feature extraction techniques, using different classification algorithms we could classify text at acceptable accuracy rate. K. Pal et al, [8] have worked on surveying the work done and challenges in managing big data. Sang-Woon K et al, have combined approach of feature weighing technique and Linear Discriminant Analysis for classification of Research Papers [9] they are introducing Big data and Machine Learning together for the classification problem. Bipanjyot K. et al. [10] have surveyed Document classification using various algorithms and performance measure used and claims there is a need to evaluate text classification using maximum evaluation metrics. Upendra S. et al. [11] have performed a survey on various classification algorithms on the basis of time complexity and performance and claims performance also varies on data collection. They also draw attention towards SVM algorithm having potential for being a good classifier but also states that universal acceptance of this algorithm is unlikely.

III. THE MODEL FOR DOCUMENT CLASSIFICATION

The model requires past experiences for learning, meaning labeled documents in this scenario. But we just should not flood with entire labeled documents at once to the model; we need to provide them the reasons based on which the document is labeled as some class. For instance a newspaper article can belong to Entertainment, Sports, Politics, national news, international news, advertisements, etc. this depends on the type of content of the news article. So we need to analyze the content and extract features to create feature set to train or provide experience to the model, Meaning before moving to Machine Learning we need to process our documents with Natural Language Processing. So we need to follow some algorithmic steps using both Natural Language Processing and Machine Learning techniques to build the final Model. The model can be broadly seen as 1. Pre - Processing Module, and 2. Feature extraction, Training and Testing Module using Machine learning techniques.

1. Pre – Processing Module

The steps followed in Pre - Processing in sequence are shown in figure 2 and listed as:

Tokenization: Separate all words separated by space and tab deli-meter, remove blank lines and generate tokens.

Data Cleaning: Remove all numbers, special characters from set of tokens

Remove Stop words: removing general and very common words, which are not contributing anything.

Stemming: Converts all words in their root form.

The result is only tokens, which can be useful for feature extraction to create Feature Set.

2. Feature Extraction, Training and Testing Module

The results generated with Pre - Processing module are used for feature extraction using Machine Learning Techniques. Features can be extracted by assigning weight to important feature by observing it occurrence using ML techniques. Once feature set is ready, it is used to train model.

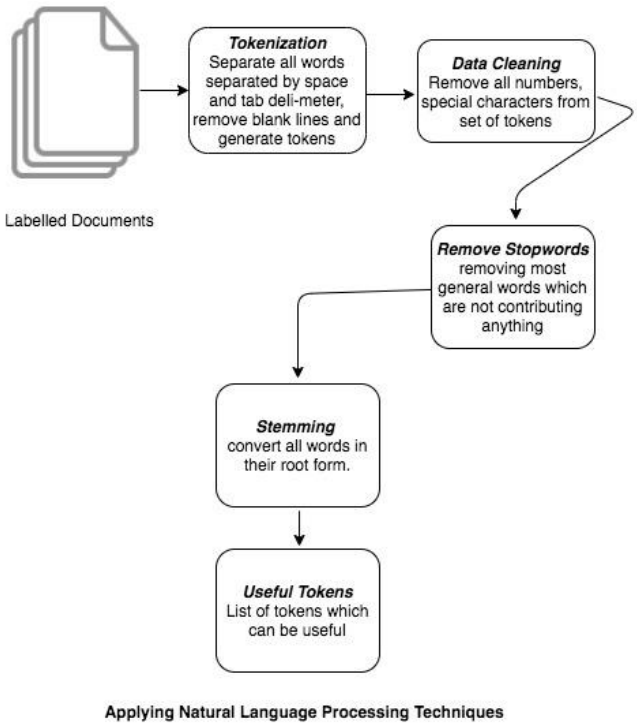

Figure 2. Pre - Processing Module

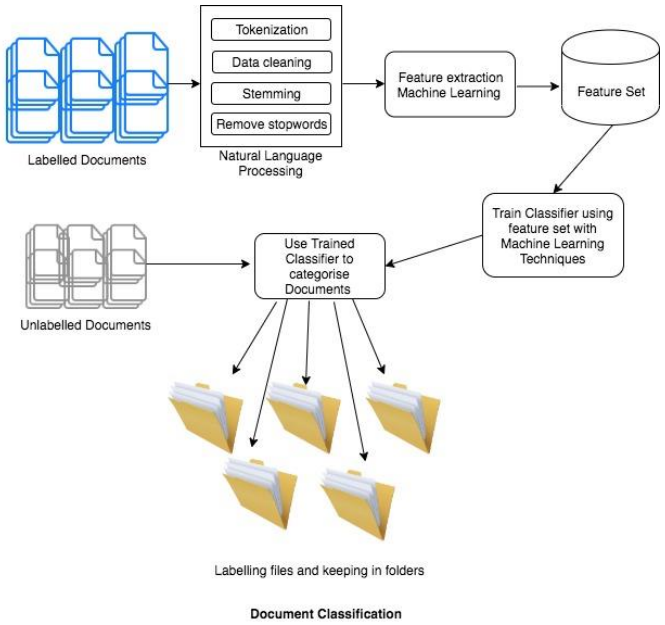

Figure 3. Architecture of System: Categorizing and Organizing Documents

The model is created using Support Vector Machine, K-nearest neighbor and Naïve Bayes Machine Learning algorithms. The trained model can predict the class of any new document resulting in placing

that file in correct location for easy retrieval and efficient usage.

The entire System along with pre-processing module based on natural language processing is shown in figure 3.

IV. EXPERIMENT AND RESULTS

The model is tested with English Movie data review with 2 categories, the review is positive or negative depending on comments provided by user. It is implemented in Python 3.6.The model has processed 2000 documents and used it to split training testing data. The Machine is trained with labeled documents, which is training data and tested with testing data. In this experiment the data is split in 4 different proportions and the results are shown in table 1. The comparative results of 50%, 60%, 70% and 80% training data for K-nearest Neighbor, Naïve Bayes and Support Vector Machine is shown in figure 4 using bar chart. Scatter plot in figure 5 is showing the accuracy with the proportion of data used fro training and how the accuracy is increasing by changing machine learning algorithm and training sample size. The accuracy is calculated by comparing real label with predicted labels of the documents.

TABLE I

THE RESULTS OF THE MODEL TRAINED WITH 4 DIFFERENT SPLITS USING 2000 DOCUMENTS

Number of Training Documents (Training Samples in %)	Number of Testing Documents (Testing Samples in %)	Name of Algorithm	Accuracy in Percentage
1000 (50%)	1000 (50%)	K- Nearest Neighbors	60.30%
		Naïve Bayes	71.50%
		Support Vector Machine	80.10%
1200 (60%)	800 (40%)	K- Nearest Neighbors	60.75%
		Naïve Bayes	70.87%
		Support Vector Machine	82.12%
1400 (70%)	600 (30%)	K- Nearest Neighbors	60.00%
		Naïve Bayes	72.50%
		Support Vector Machine	83.00%
1600 (80%)	400 (20%)	K- Nearest Neighbors	65.75%
		Naïve Bayes	74.50%
		Support Vector Machine	85.00%

The K nearest neighbor is used with 7 neighbors, Support Vector Machine is used 'rbf' kernel and Gaussian Naïve Bayes is used for this experiment.

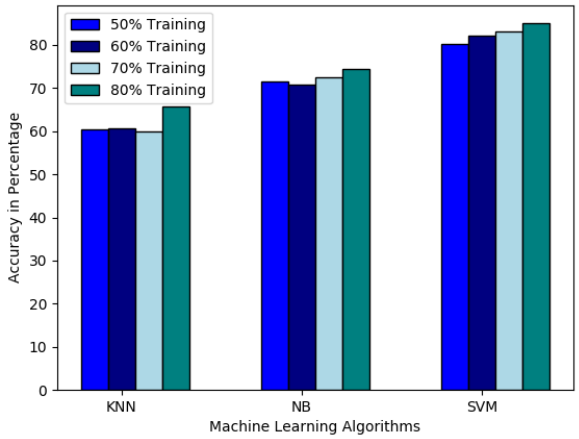

Figure 4. Comparative results of 3 Algorithms with 50%, 60%, 70% and 80% Training samples

Figure 5. Scatter results of 3 Algorithms with 50%, 60%, 70% and 80% Training samples

V. CONCLUSION

The model is developed using Natural Language Processing techniques and Machine Learning techniques and is tested with English Movie data review with 2 categories. The model used 3 different Machine Learning algorithms for classification of documents namely K-nearest Neighbor, Naïve Bayes and Support Vector Machine and trained with 50%, 60%, 70% and 80% documents and tested with 50%, 40%, 30% and 20% of documents respectively. The experiment shows increasing training examples increases the performance consistently for all algorithms used. Support Vector Machine is performing well with 80.10%, 82.12%, 83.00% and 85.00% accuracy with minimum to maximum training samples. Naïve Bayes is second in terms of performance with more than 70% accuracy and KNN is performing the least with accuracy ranging from 60% to 65.75%. The model is tested with English Documents, but it can be applied to any language supported by machine but will have different accuracy. To see the effective of the model it needs to be tested with documents of other languages and accuracy can be improved with different feature extraction, feature Selection techniques and using other sophisticated algorithms available in machine learning.

VI. REFERENCES

[1] Jayashri K., Mayura K. (2013) "Machine Learning Algorithms for Opinion Mining and Sentiment Classification" International Journal of Scientific and Research Publications, Volume 3, Issue 6. 724 - 729.

[2] Kaur, Jasleen and Jatinderkumar R. Saini.(2017) "Punjabi Poetry Classification: The Test of 10 Machine Learning Algorithms." International Conference on Machine Learning and Computing (ICMLC 2017)-ACM

[3] Harikrishnna D M, K. Sreenivasa Rao. (2015) Children Story Classification based on Structure of the Story. IEEE International Conference on Advances in Computing, Communications and Informatics. 1485-1490

[4] K. Pal, B. V. Patel (2020). "Model for Classification of Poems in Hindi Language Based on Ras", Smart Systems and IoT: Innovations in Computing, Smart Innovation, Systems and Technologies, Springer, 141. 655 – 662

[5] Shalini Puri, Satya Prakash Singh, (2018). "Hindi Text Document Classification System Using SVM and Fuzzy: A Survey", International Journal of Rough Sets and Data Analysis.

[6] Shalini Puri, Satya Prakash Singh.(2019) An Efficient Hindi Text Classification Model Using SVM Computing and Network Sustainability Book.

[7] K Pal, B V Patel, (2017) "A Study of Current State of Work Done for Classification in Indian Languages", International Journal of Scientific Research in Science and Technology. 3(7) 403 – 407

[8] K Pal, J Saini, (2014) "A study of current state of work and challenges in mining big data", International Journal of Advanced Networking Applications. 73 – 76.

[9] Sang-Woon Kim1 and Joon-Min Gil (2019), "Research paper classification systems based on TF-IDF and LDA schemes", Human- Centric Computing and Information Science, 1 – 21.

[10] Bipanjyot Kaur , Gourav Bathla (2018), "Document Classification using Various Classification Algorithms: A Survey", International Journal on Future Revolution in Computer Science & Communication Engineering. 4(2), 150-155

[11] Upendra Singh, Saqib Hasan (2015), "Survey Paper on Document Classification and Classifiers", International Journal of Computer Science Trends and Technology, 3(2) 83 – 87.

Cite this article as :

Kaushika Pal, Dr. Biraj V. Patel, "Multi - Class Document Classification: Effective and Systematized Method to Categorize Documents ", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 7 Issue 7, pp. 118-123, January-February 2020. Available at doi : https://doi.org/10.32628/IJSRSET207117
Journal URL : http://ijsrset.com/IJSRSET207117