IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 6, June 2021

DOI 10.17148/IJARCCE.2021.10694

Text Document Classification Using Machine Learning Techniques

Sakshi Ghodke[1], Suvarna Gavai[2], Shubhada Gaikwad[3], Gayatri Inamdar[4], Prof. V.S. Kolekar[5]

TSSM's Bhivarabai Sawant College of Engineering and Research, Narhe, Pune-411041

Abstract – The association classification technology based on frequent patterns is recently presented, which build the organization rules by frequent patterns in various categories and classify the new text employing these rules. However, in the current association organization methods, shortage exists in two aspects when it is applied to classify text data: one is the method ignored the information about word's occurrence in a text; the other is, the method needs pruning rules when the mass rules are generated, but that leads the veracity of classifying to drop. Therefore, this paper presents a text classification algorithm based on frequent pattern with term frequency, and obtains higher performance than other association categorization methods and some current text classification methods. Our study provides suggestion that association rule mining can be used for the construction of fast and effective classifiers for automatic text categorization.

Key Words: Machine Learning, CNN Algorithm.

1. INTRODUCTION

Automated feature selection is important for text classification to reduce the feature size and to speed up the learning process of classifiers. In this paper, we present a novel and well-organized feature selection framework based on the Information Theory, which aims to rank the features with their discriminative capacity for classification. Disadvantages: Accuracy is very low based on word count. It can be improved by using any other classification algorithms like Naïve Bayes. Text categorization is a process in data mining which assigns predefined groups to free-text documents using machine learning techniques. Any file in the form of text, image, music, etc. can be classified using some categorization techniques. . Disadvantages: Accuracy is not compared with much data. In data mining the more the data, proper results can be found.

Objective :-

We have to perform text reduction. Generate Frequencies based on Modified algorithm, Classification of unknown input using proposed algorithm, comparison of existing and proposed system.

1.1 Methodology:-

This system what does is that it takes in unknown inputs from the user which needs to be classified for them. The Text classifier algorithm i.e. the CNN algorithm of text classifier compares the input with the dataset and then using the machine learning classifies the documents.

2. SYSTEM ARCHITECTURE: -

The lower figure gives you the complete idea of several connections recognized between the microcontroller and extra sensors for decent functionality

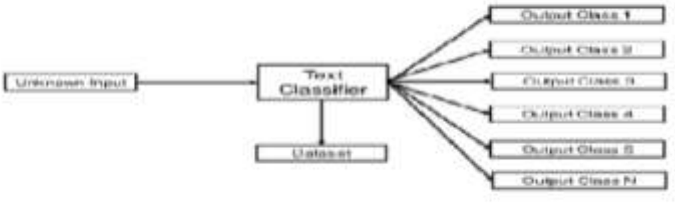


Fig: Working of Text Classifier

3.1 Block diagram of The System

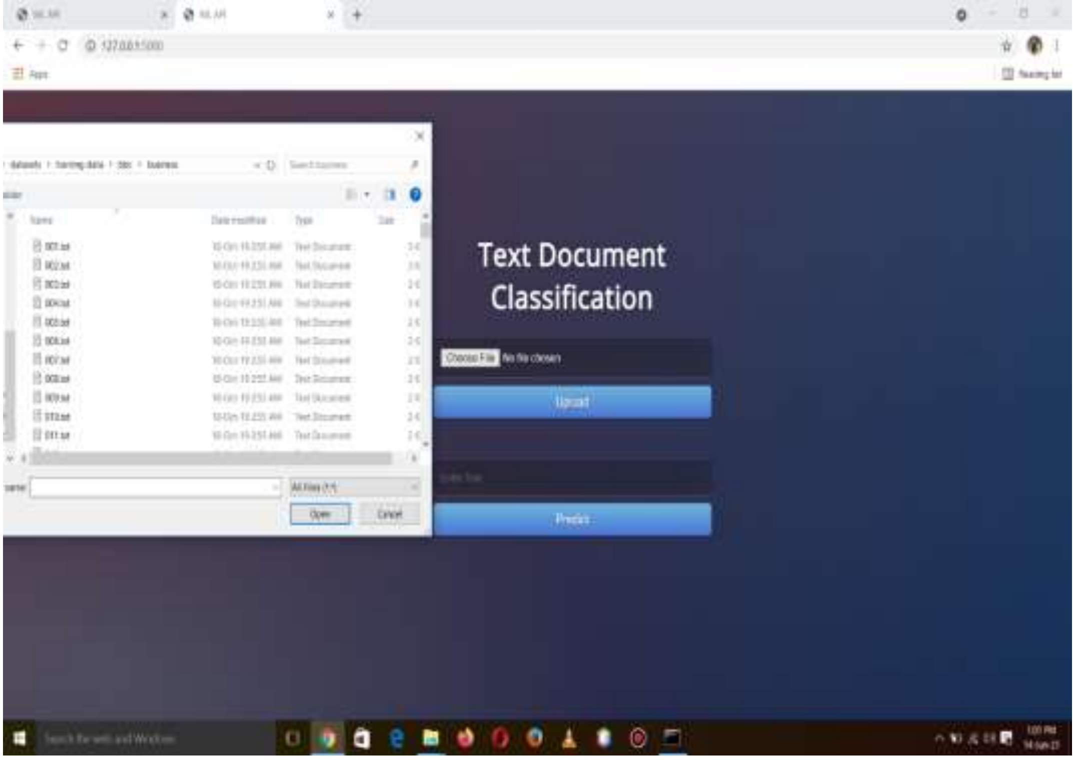
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

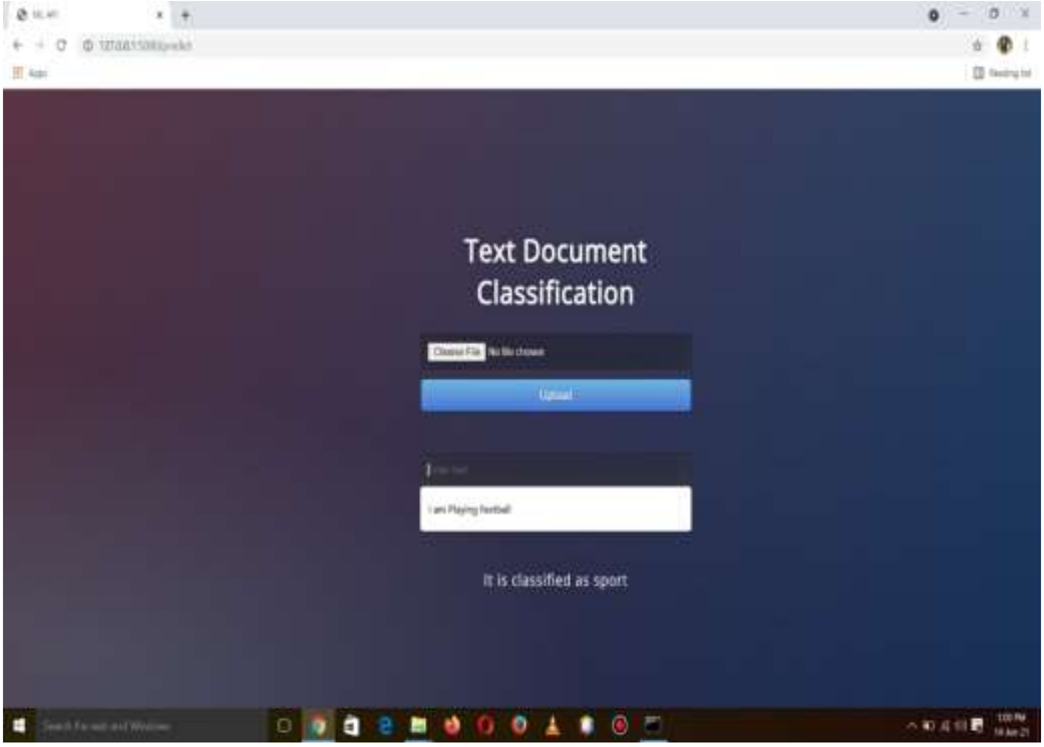
International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 6, June 2021

DOI 10.17148/IJARCCE.2021.10694

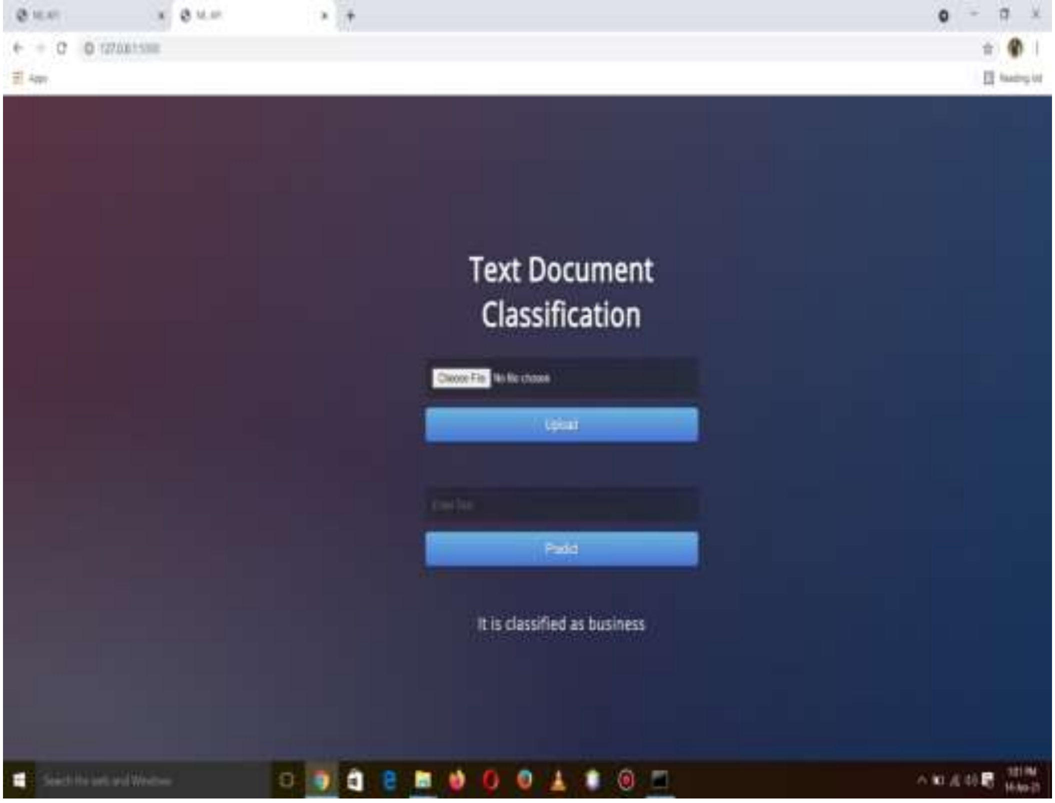
3. EXPERIMENTAL RESEARCH-





I am Playing football

It is classified as sport



It is classified as business

4. ALGORITHM

Step 1- we are going to give the whole document as an input to our model. To Process these documents, it must be first converted to a suitable format for processing. Text Documents are theeasiest format to read and process, so first step is to convert the document into text format and thenextract the text into list in python. The classification model was built using Convolutional NeuralNetwork, using keras. Keras is a library in neural networks and provides APIs for Neural Network, which is included in tensorflow which is an open source library for machine learning projects. Theconvolution layer applies filters of different size to these word embedding matrix to extract featuresas vector corresponding to each filter. We are using convolution filter of sizes 3,4,5 in our model.

Step 2- involves, data Preparation and pre-processing: The preprocessing done on the data set are:remove punctuations, convert to lowercase, remove numbers and remove single characters.Converting texts into lowercase helps in parsing the text better it also helps in reducing the size ofthe vocabulary. Removing punctuations helps to increase the coverage of the embeddings on our vocab. Removing numbers and removing single characters helps in improving the embedding ratio by 2-3%.The input is then passed to the embedding layer. Embeddings are representation of words into a type such that the words with similar meaning have similar representation. This contracts to the thousands or millions of dimensions for the representation of sparse words like one hot encoding.The input to Convolutional Neural Network contains categorical features that takes one of k distinct symbols, each word is associated to each possible feature value with a dimension d. Embeddings helps to capture the implicit relations between words, by finding the number of timethe word is used in the training documents. In our model we are using glove (Global Vector for Word Representation). It is a method for efficiently learning word vectors. Glove is an unsupervised learning algorithm for generating the vector representation for the words. The training if performed on the word-word co-occurrence statistics from the corpus, and the representation depicts a linear substructure of the words. The output of each convolution layer will go through a maxpooling layer. Maxpooling method helps to combine the vectors from different convolution filters into a single-dimensional vector. This is done by taking the max value observed in resulting vector from the convolutions. Max pooling chooses the maximum of the value from the input feature map. This is done to reduce thenumber of vectors in the feature map. This feature is then passed to the softmax activation layer that converts the output into the probabilities that result into 1. The values given to the fully connected neural network layer, CNN goes through the back propagation process to determine themost accurate weights. The model with the highest accuracy is saved in the hierarchical data formatand the file is updated if the accuracy of the model increases from the previous value.

DOI 10.17148/IJARCCE.2021.10694

5. FUTURE SCOPE

For the future scope of the model we can add more classes to the model or we can make it a generalize model which will be able to classify the text in most of the classes and detect it. We canalso add the image classification method to the model for detection of the classes using the imagedata set for testing as well as training dataset. Also further the model can also be built as to more specialized categories such as the files belongs to a particular class such as cricket or football on sports categories.

6. CONCLUSION

Document Classification using neural networks has a better analytic results as compared with othertraditional classification methods. The use of word embedding for finding the features of the documents helps to classifier to get the most frequently used words and its relation with other words. This helps the classification to get the context of the word in text document. Neural Network also helps in finding the relations between the words due to its ability to find the correlation between the words. Due to this neural networks can also find relationships on unseen data as well, so it can classify data effectively using these correlations.

ACKNOWLEDGEMENT

With immense pleasure, I am presenting this project " Text Document Classification Using Machine Learning Techniques " as part of curriculum of B.E. Computer Engineering. I wish to thank all people who gave me unending support right from the stage the idea was conceived. I am thankful to my Guide Prof. V. S. Kolekar her great support throughout the course of this project. I also thank all those who have directly or indirectly guided and helped in the preparation of this project. I express a profound thanks to our respected Head Of The Department Dr.S.U.Kadam who advise and valuable guidance helped in making this seminar successful.

REFERENCES:

[1]Sang-Bum Kim, “Some Effective Techniques for Naive Bayes Text Classification” PhD, Department of Computer Science, Korea University, 2006.

[2]Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) Convolutional Neural Networks for Sentence Classification by Yoon Kim.

[3]Dino Isa, Lam Hong Lee, “Text Document Pre-processing with the Bayes Formula for Classification Using the Support Vector Machine”, IEEE TRANSACTIONS on data engineering2008

[4]Nihar Ranjan, Rajesh Prasad, “Automatic Text Classification using BP Lion- Neural Networkand Semantic Word Processing”, Imaging Science Journal Print ISSN: 1368-2199, Online ISSN: 1743-131X

[5]D.D. Lewis, “Representation and Learning in Information Retrieval,” PhD dissertation, Dept. of Computer Science, Univ. Of Massachusetts, Amherst, 1992.

[6]Jung-Yi Jiang, Ren-Jia Liou, “A Fuzzy Self-Constructing Feature Clustering Algorithm for Text Classification”, IEEE Transactions on Data Science 2011.

[7]E. H. Huang, R. Socher, C. D. Manning, and A. Y. Ng, ``Improving word representations via global context and multiple word prototypes," in Proc. ACL, Jeju-do, South Korea, Jul. 2012, pp.873_882.