Multi Label Classification Of News Text

Can Koral ADALI - Miray PADIR – Bihter ÖZUÇAK
Kardelen YILDIRIM
Erdoğan DOĞDU – Roya CHOUPANI

Çankaya University, Department of Computer Engineering

Abstract

By the development of technology, there are a lot of information on the internet and day-by-day there is huge amount of increase in number. Every information/data needs to be classified by the subject of them. Our research covers what are the types of classification, how to classify them using the multi-label classification algorithms and usage cases of classification. In this paper, we study and mention that we're going to use word embedding with word2vec as an algorithm for classifying news text as multi-labels. Moreover, we are going to build a web-interface with PHP so that, user can upload a news text and get the result of it as categories. We identify and state that there are many algorithms and methods can be used for classification.

Keywords: Multi-label classification, Multi-class classification classification, word embedding, word2vec, python

Introduction

Nowadays, there are a lot of informations on Internet and every information has their own classification which can be related with medical, marketing, news and many more. Every information has and needs some kind of classification and it needs to be classified for quick access and preventing the data loss. The classification has been widely studied and it has more than one way to classify an information with computer which can be generalizable by one main topic which is Machine-learning.This study focuses on news text multilabel classification. News have different subjects and some news can have more than one subject that needs to be classified. So that every news text needs multilabel classification. For our study, we choose using Python language which is mainly used for deep learning. We decided to use Word Embedding with lda2vec which can be done with Python and build a web-interface with PHP which helps user to upload a news text and gets the results.

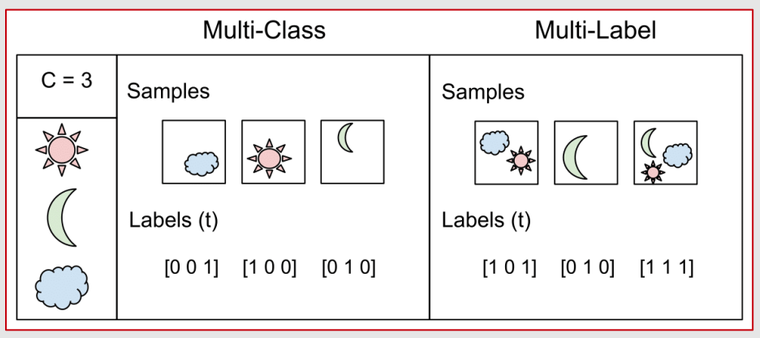

Figure 3 – Multi Class And Multi Label

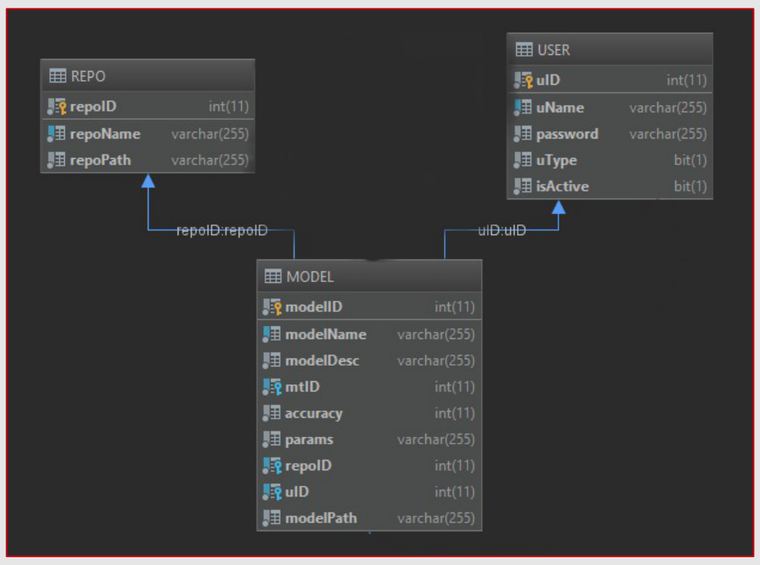

Figure 1 - Database

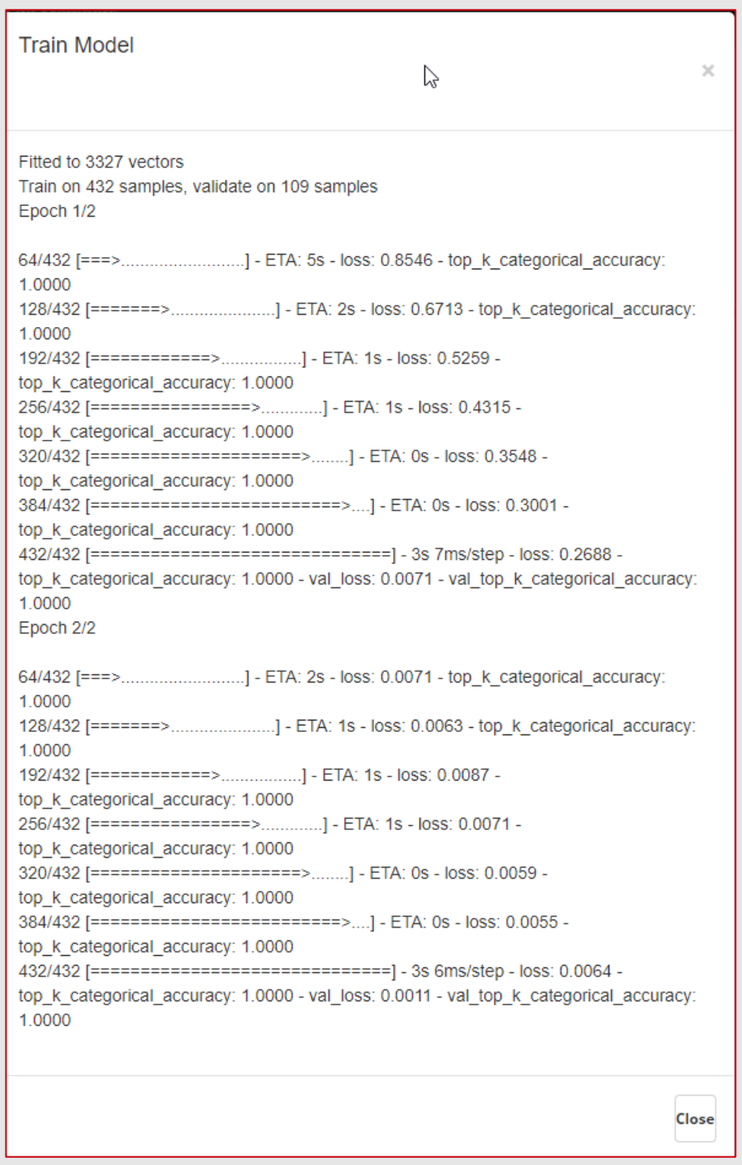

Figure 2 – Train Model

Solution

Our solution based on two processes. One of them is train model based on given datasets. The other one is using this trained models on given news text documents. In detail, we accomplished the training task by using deep learning. Training is performed by using Magpie library which is using Google's, Tensorflow and Keras modelling algorithm.Finally we used this technique for our modelling and then used PHP for Web Application to interpret giving news text.

Results & Conclusion

In this project, a web application developed which aims to easily determine the multiple texts of Turkish news texts with a classification model formed from datasets with millions of news. This application aims to make it more informed and effective for the best user experience in the use of news sites. As a result of the classification, the application will determine the topics according to the relevance of the text. As a result of this percentage, the user can publish the news according to the level of interest of the classified news under the headings.

Acknowledgement

We are greateful for guidance we have received from Prof. Dr. Erdoğan DOĞDU and Assist. Prof. Dr. Roya CHOUPANI. The help we recieved from them was a great asset to improve this project and ourselves.