HIGH LEVEL DOCUMENT

News Classification Application

Written By	**Mohammad Sohail Parvez**
Document Version	**0.1**
Last Revised Date	

Content

Abstract

For this project, we examine the “news-classification” dataset available at the kaggle.We aim to classify the news according to the category(Business, Sports, Technology, Entertainment, Politics)

1. Introduction

1.1 Why this High-Level Design Document?

The purpose of this High-Level Design (HLD) Document is to add the important details about this project. Through this HLD Document, I'm going to describe every small and big thing about this project.

2.General Description

2.1 Product Perspective

The News Classification using Supervised learning(classification) based on Natural Language Processing with Machine Learning.

2.2 Problem statement

In today's world, data is power. With News companies having terabytes of data stored in servers, everyone is in the quest to discover insights that add value to the organization. With various examples to quote in which analytics is being used to drive actions, one that stands out is news article classification. Nowadays on the Internet there are a lot of sources that generate immense amounts of daily news. In addition, the demand for information by users has been growing continuously, so it is crucial that the news is classified to allow users to access the information of interest quickly and effectively. This way, the machine learning model for automated news classification could be used to identify topics of untracked news and/or make individual suggestions based on the user's prior interests.

2.3 Proposed Solution

The dataset contains a lot of news details which are raw data. The dataset has many punctuations, stopwords, and the words are capitalized so we have to lower them. After preprocessing our raw data, we are going to encoding the labels to numeric.Them pass to different deep learning algorithms(Baseline, Simple Dense, Conv1D, Bidirectional, LSTM, GRU)

2.4 Technical Requirements

In this project the requirements to get income classify various platforms. For that, in this project we are going to use different technologies. Here are some requirements for this project.

- Model should be exposed through API or User Interface, so that anyone can test model.
- Model should be deployed on the cloud..

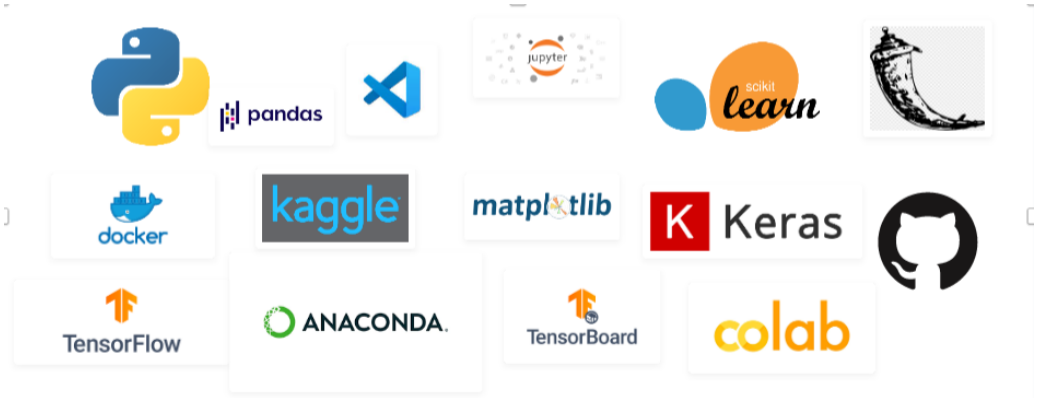
2.5 Data Requirements

The dataset is downloaded from Kaggle. The dataset contains 1400+ records for training and 700+ for testing data.

The train dataset contain three columns

- **ArticleID**: Unique Id for each news
- **Text**: Particular text of the header and article.
- **Category**: Category of the article(Tech, Business, Sport, Entertainment, politics)

2.6 Tools Used

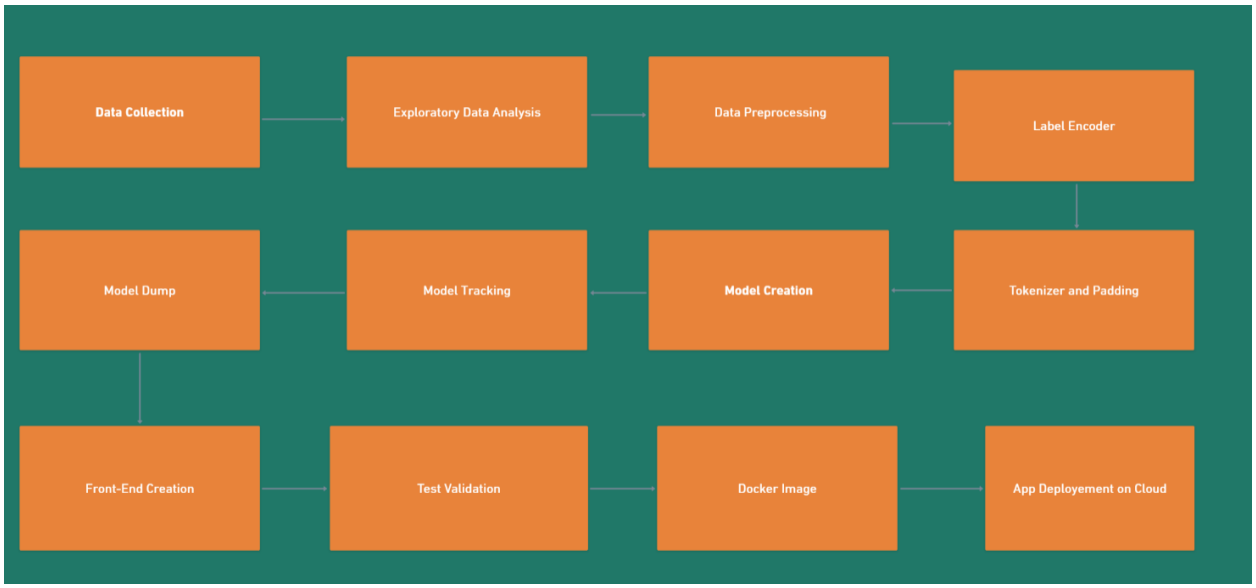


2.7 Constraints

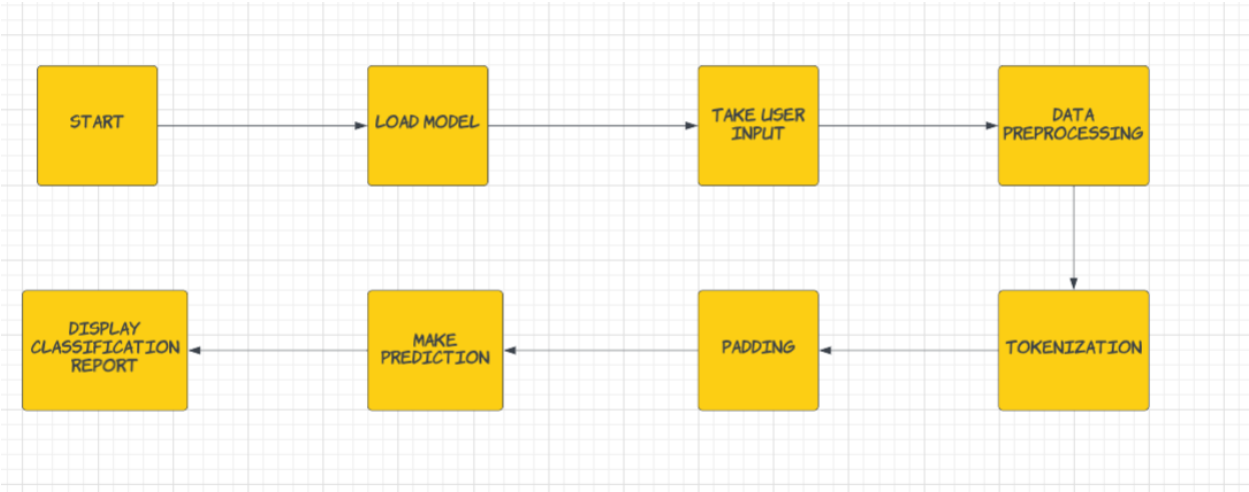
The news classification prediction system must be user friendly, error free and users should not be required to know any of the back-end working.

3 Design Details

3.1 Process Flow



3.2 Deployment Process



3.3 Event Log

In this project we are logging every process so that the user will know what process is running internally.

Step-By-Step Description:

- In this project we defined logging for every function.
- By logging every function in the Exploratory Data Analysis.
- Then logs all the data preprocessing functions and code.
- Logs every model algorithm to our data.

3.4 Error Handling

The project is designed in such a way that, at any step if error occurs then our application should not terminate rather it should catch the error and display that error with proper explanation as to what went wrong during process flow.

4. Performance

Solution of News Classification is used to classify in advance, so it should be as accurate a s possible so that it should give as much as possible accurate classification. That's why before building this model we followed the complete process of NLP with Machine Learning. Here are summary of complete process:

1. First we cleaned our dataset properly by removing all null value and duplicate values present in the dataset

2. Then we explored our dataset using visualization. We can see that there is a proper distribution of counts in different categories of news(Business, Tech, Politics, Sports, Entertainment).

3. Data preprocessing by lowering the sentences, removing punctuations, numbers, character, multiple spaces using regex.

4. Data splitted into independent and dependent features. After independent and dependent features it has further splitted into training and validation datasets.

5. Before we start model training we have to handle the categorical labels. To do that we have labeled it using scikit-learn framework with one-hot and label encoder.

6. In model training we used six different models. Baseline, Simple_dinse, Conv1D, LSTM, GRU, Bidirectional LSTM. In these six models Conv1D outperformed the rest of the models.

7. Save the 'Conv1D' model.

8. Let's test our model on unseen data (Test dataset).

9. After the model was ready to deploy. We deployed on the cloud.

4.1 Re-usability

The programming is done in such a way for this project that it should be reusable. So that anyone can add and contribute without facing any problems.

4.2 Application Compatibility

The difference module of this project is using Python as an interface between them. Each module has its own job to perform and it is the job of the Python to ensure the proper transfer of information.

4.3 Deployment

We have deployed this on cloud and also dockerized this.

5. Conclusion

The News Classification model will classify whether the given news is in which category.