High Level Document (HLD)

News Article Sorting

- Rajat Chaudhari

Contents

1. Introduction

This project is a NLP project which helps to classify news articles into one of the 5 categories – business, entertainment, politics, sport, technology. The purpose of this High Level Document is to add the necessary details to the current project description to represent a suitable model for coding. This document includes a high-level architecture diagram depicting the structure of the system, such as the hardware, database architecture, application flow, technology architecture.

2. Scope

This project will produce a deep learning classification model that will be able to classify news articles into of the five categories (business, entertainment, politics, sport, technology).

The project will provide a UI to be able to use the deep learning model. The UI will provide following information:

(i) Level of confidence in prediction of each category
(ii) Final prediction for the news article input

One of the major limitations of this project is the small size of the dataset being used to train the model.

3. 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.

4. Proposed Solution

This project proposes a solution in the following way:

(i) Train a deep learning model to train on the provided dataset. The model should be able to classify news articles with a high level of accuracy.

(ii) A UI is provided for a user to use the deep learning model.

5. Further Improvements

The proposed solution can be deployed on a cloud. Devops principles and pipelines can be used to develop to further develop the project. Pipeline to automate monitoring, or automate update integration to the project can be implemented.

6. Data and Database

DataStax Astra (Cassandra) database is used to store the data. The data consists 1490 records for training set, 736 records for test set. The data fields in the dataset are:

(i) ArticleID
(ii) Article
(iii) Category

7. Tech Stack

- Streamlit
- Python
- Tensorflow
- Keras
- Nltk
- Jupyter Notebook

Streamlit – to deploy the machine learning model with a user interface (UI) on local host

Python – to develop the overall project

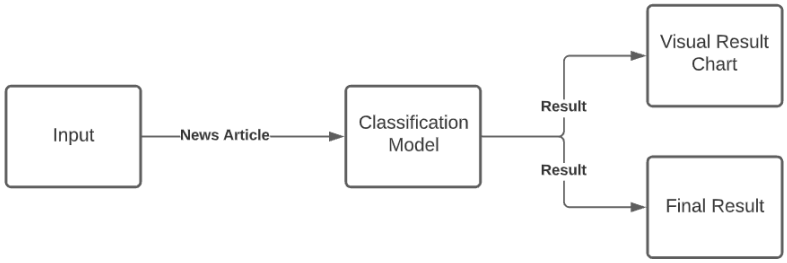
Tensorflow, Keras – framework to develop the deep learning model

NLTK – to process text data

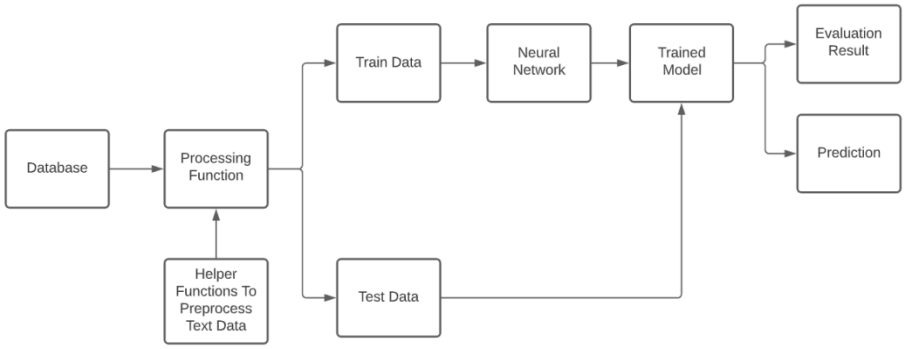
Jupyter Notebook – environment/ tool to develop the deep learning model

8. Design Details

a. Process Flow



b. Model Training



9. Logging

Python library logging is used to log different kinds of messages to a log file.
There are 5 levels of logging messages in the logging python library, namely:

1. DEBUG
2. INFO
3. WARNING
4. ERROR
5. CRITICAL

For this project, the level is set to INFO. All the logging messages are written to logs.txt file.

10. Conclusion

This project provides a User Interface to a deep learning classifier model. The project is capable of accepting a news article input from the user in the form of a text. The model then processes the article and then generates a bar chart graph. The bar chart graph displays the percentage of which category the model thinks the news article belongs to. The system also displays a final result classifying the news article in one of the 5 categories (business, entertainment, politics, sport, technology).