Suraj Verma

High Level Design (HLD)

News Articles Sorting

Document Version Control

Date Issued	Version	Description	Author
21/03/23	1	Initial HLD – V1	Suraj Verma
31/03/23	2	Updated Performance, Conclusion and References	Suraj Verma

Contents

Abstract

As the world progresses with advanced technologies and comforting appliances With the presence of textual data everywhere in the internet, it becomes necessary to structure this data for profitable analysis and exploitations. It is necessary to categorize data with respect to its domain for the appropriate arrangement of the data.

News Article Classification is a process of automatically categorizing news articles into predefined categories based on their content. It is an important task for news organizations to organize their content and make it easily accessible to users. The process typically involves the use of machine learning algorithms to analyze various features of the articles, including text, images, and metadata, and assign them to relevant categories such as sports, politics, entertainment, and technology. The accuracy of the classification depends on the quality of the algorithms, the size and diversity of the training data, and the effectiveness of feature selection techniques. The task of news article classification is ongoing and requires continuous improvement and testing to ensure high accuracy and relevance for users.

1 Introduction

1.1 Why this High-Level Design Document?

The purpose of this High-Level Design (HLD) Document is to add the necessary detail to the current project description to represent a suitable model for coding. This document is also intended to help detect contradictions before coding and can be used as a reference manual for how the modules interact at a high level.

The HLD will:

- Present all of the design aspects and define them in detail
- Describe the user interface being implemented
- Describe the hardware and software interfaces
- Describe the performance requirements
- Include design features and the architecture of the project
- List and describe the non-functional attributes like:

  -Security
  -Reliability
  -Maintainability
  -Portability
  -Reusability
  -Application compatibility
  -Resource utilization
  -Serviceability

1.2 Scope

The HLD documentation presents the structure of the system, such as the database architecture, application architecture (layers), application flow (Navigation), and technology architecture. The HLD uses non-technical to mildly-technical terms which should be understandable to the administrators of the system.

2 General Description

2.1 Product Perspective

The goal of this project is to build a model that can predict the appropriate category for a news article. For this, We have used Natural Language Processing algorithms to develop the prediction model.

2.2 Problem Statement

To create news classification model by using machine learning algorithms which will help us to predict the appropriate domain of a new article.

2.3 Proposed Solution

The solution proposed here is a prediction model which will help us to find out the correct domain of a given news article using Natural Language Processing. This entire project is build using DVC (Data Version Control) and MLFlow for retraining, experimenting and model monitoring purpose.

2.4 Further Improvements

With continuous retraining and model configurations, model can be improved continuously. In future, other features can be added to improvise our prediction model.

2.5 Data Information

The dataset used in this project is the BBC News Article Classification dataset . The data is available in Kaggle Datasets.
Every transformation of the data is uploaded to and extracted from MongoDB database. The dataset contains three fields, Ariticle ID, Text and Category.

Information of attributes of dataset as follows:

- ArticleID : The Unique ID of the article.
- Text : Text Article.
- Category : The Relative domain of the given text article.

2.6 Tools Used

The whole project is built upon Python Programming language. Visual Studio code and jupyter notebooks are used as a prime development interface. Python libraries such as NumPy, Pandas, Matplotlib, DVC, MLflow, etc. are majorly used in the project.

- Visual Studio Code is used as IDE.
- Python is used as the main programming langiuage.
- HTML is used for website design.
- Git is used for continous intergration and version control.
- DVC is used for Data Version Controls
- MLflow is a platform to streamline machine learning development, including tracking experiments, packaging code into reproducible runs, and sharing and deploying models.
- Jupyter is used for analysis notebooks and for testing codes.
- Sklearn is used for performing machine learning operations.
- Python libraries such as numpy, pandas, matplotlib, seaborn, etc. are also used for required functions.

- Railways and Heroku are the deployment platforms used for this project.

2.7 Constraints

The machine learning based prediction system must be user friendly, as automated As possible and users should not be required to know any of the workings.

2.8 Assumptions

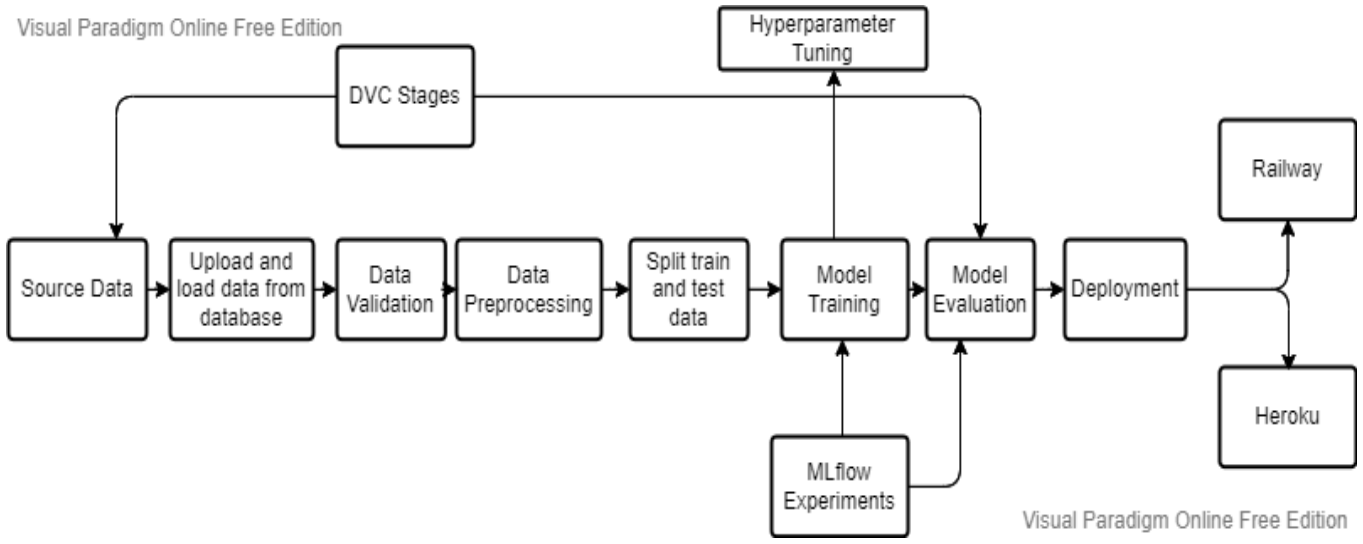
The main objective of this project is to build a predictive model using the NLP techniques such as Bag of Words, Stemming, Lemmatization, TF-IDF, etc. It is assumed that the project has to ability predict accurate results when appropriate requirements are satisfied. Model retraining can be performed with the help of mlflow.

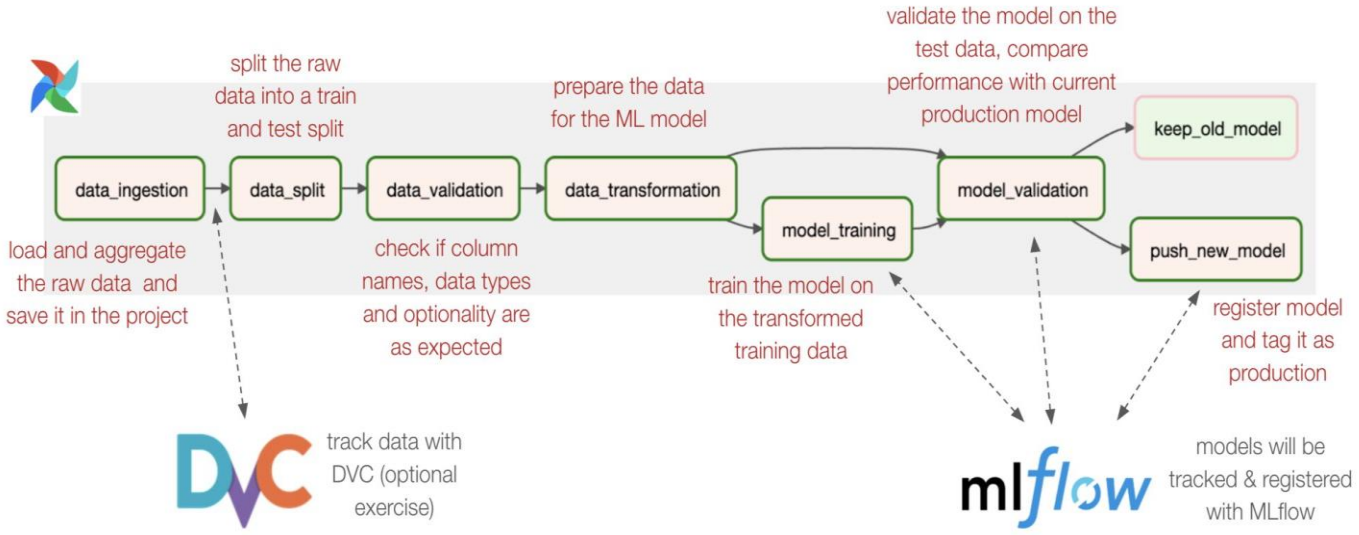
3 Design Details

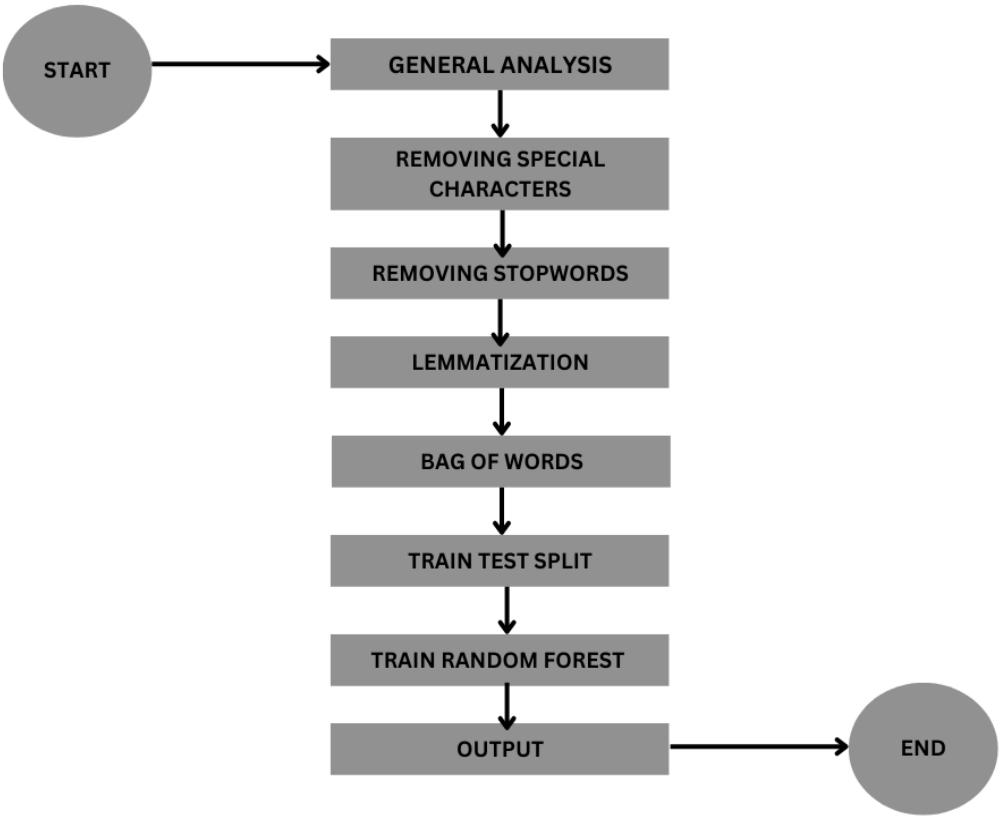
3.1 Process Flow

For prediction model, we will be using machine learning models. Below is the process flow diagram of the project.

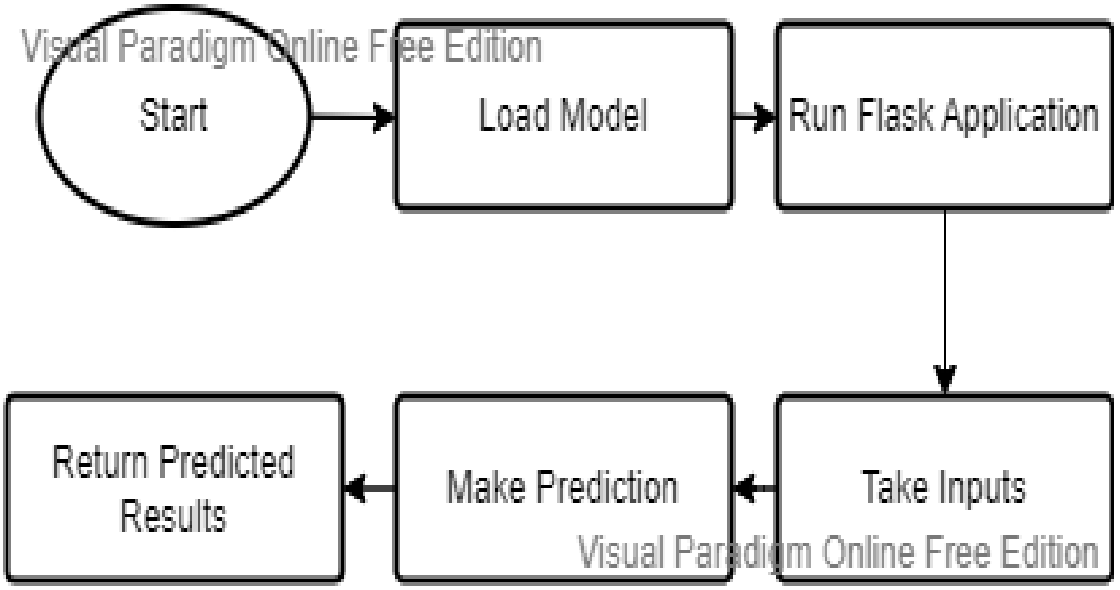
3.1.1 Proposed Methodology







3.1.2 Deployment



3.2 Event log

The system should log every event so that the user will know what process is running internally.
Initial Step-By-Step Description:
The System identifies at what step logging required
The System should be able to log each and every system flow.
Developer can choose logging method. You can choose database logging/
File logging as well.
System should not hang even after using so many loggings.
Logging just because we can easily debug issues so logging is mandatory to do.

3.3 Error Handling

Should errors be encountered, an explanation will be displayed as to what went wrong? An error will be defined as anything that falls outside the normal and indented usage.

4 Performance

The New Article Sorting project is a prediction system which predicts the category of the given article using NLP techniques. The model therefore, is expected to be accurate as overfitting or an underfitting model can lead to unexpected results.

4.1 Error Handling

The code written and the components used should have the ability to be reused with no problems. Custom Exception class is used to handle errors.

4.2 Application Compatibility

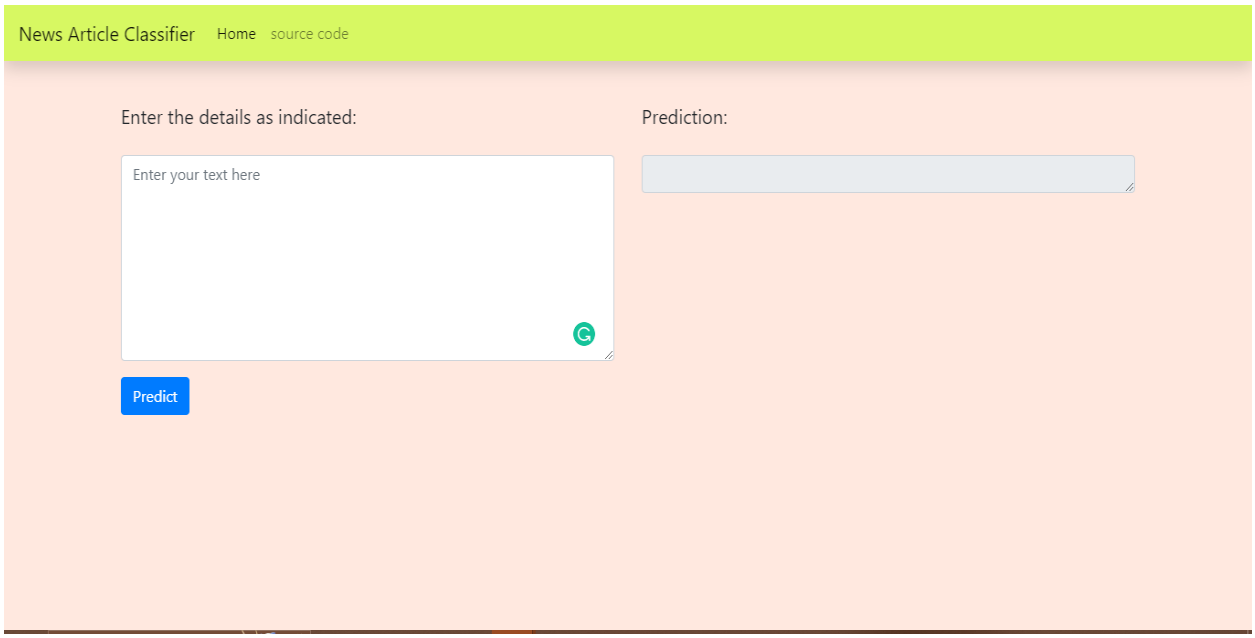
The different components for this project will be using Python as an interface between them. Each component will have its own task to perform, and it is the job of the Python to ensure proper transfer of information.

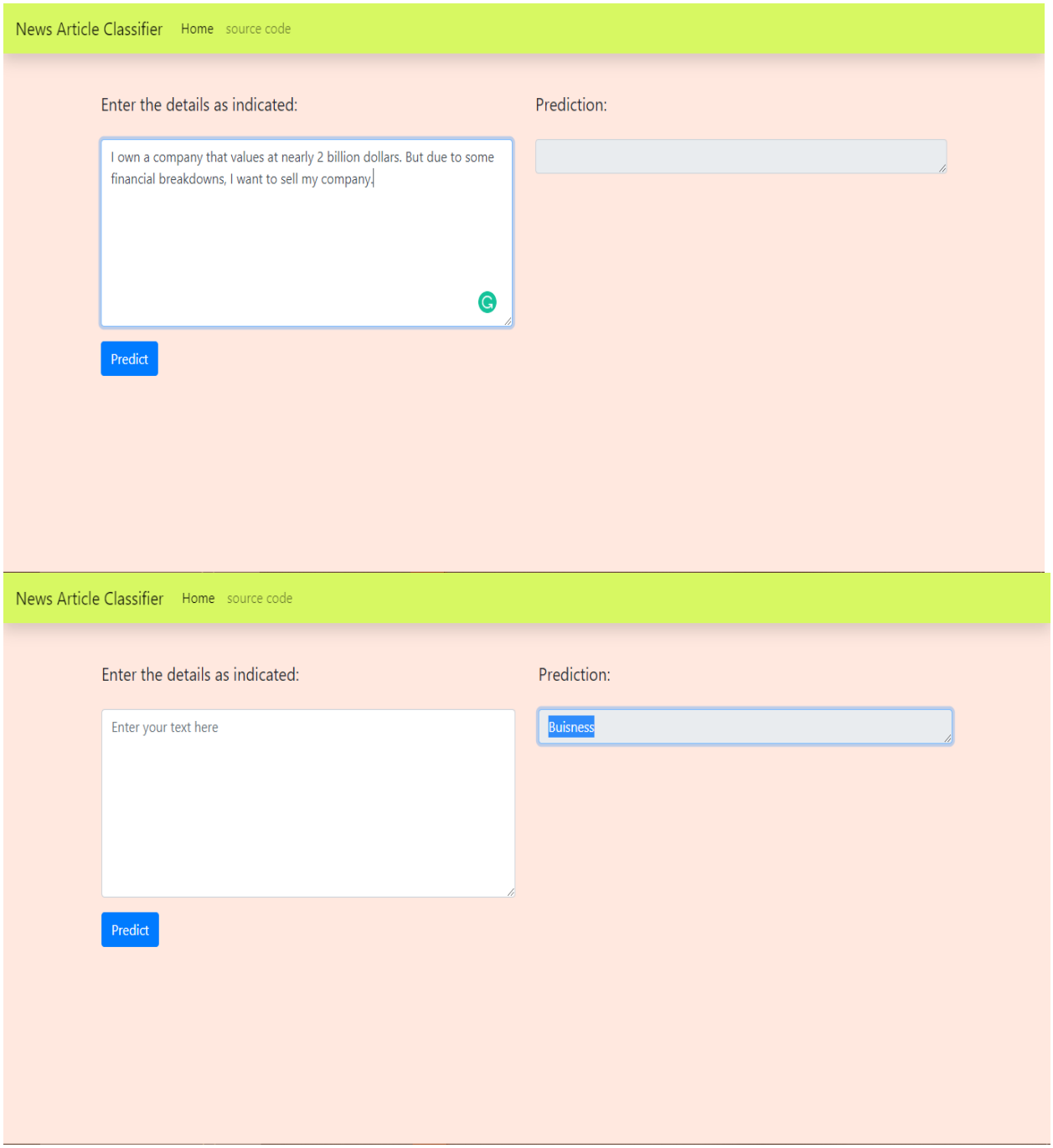
4.3 Resource Utilization

When any task is performed, it will likely use all the processing power available until that function is finished.

4.4 Deployment

- Railway - Railway is an infrastructure platform where you can provision infrastructure, develop with that infrastructure locally, and then deploy to the cloud.
- Heroku - Heroku is a platform as a service (PaaS) that enables developers to build, run, and operate applications entirely in the cloud.
- Render - Render is a unified cloud to build and run all your apps and websites with free TLS certificates, global CDN, private networks and auto deploys from Git.





5 Conclusion

The aim of this project is to build a prediction model which will help us to find out the traffic volume based on different factors such as time, weather conditions and holidays. With continuous improvisations the project can upgrade its accuracy and parameters and hence can give insights to real world traffic problems.

6 References

- https://www.kaggle.com/competitions/learn-ai-bbc/data
- https://www.youtube.com/playlist?list=PLZoTAELRMXVOk1pRcOCaG5xtXxgMalpIe