**High Level Design (HLD)**

*News Article Sorting*

*Created By-:*

*Shraddha Manoj Gujar*

## **Contents**

[Abstract. 4](#_heading=h.30j0zll)

1. [Introduction 5](#_heading=h.1fob9te)
   1. [Why this High-Level Design Document?. 5](#_heading=h.3znysh7)
   2. [Scope. 5](#_heading=h.2et92p0)
   3. [Definitions 5](#_heading=h.tyjcwt)
2. [General Description. 6](#_heading=h.1t3h5sf)
   1. [Product Perspective 6](#_heading=h.4d34og8)
   2. [Problem statement 6](#_heading=h.2s8eyo1)
   3. [PROPOSED SOLUTION 6](#_heading=h.17dp8vu)
   4. [FURTHER IMPROVEMENTS 6](#_heading=h.4i7ojhp)
   5. [Technical Requirements. 6](#_heading=h.3rdcrjn)
   6. [Data Requirements 7](#_heading=h.26in1rg)
   7. [Constraints](#_heading=h.2xcytpi) 8
   8. [Assumptions.](#_heading=h.35nkun2) 8
3. [Design Details](#_heading=h.1ksv4uv) 9
   1. [Process Flow.](#_heading=h.44sinio) 9
      1. [Model Training and Evaluation](#_heading=h.2jxsxqh) 9
      2. [Deployment Process](#_heading=h.z337ya) 10
   2. Performance 11
   3. Deployment.
   4. Future Enhancements
4. Conclusion 12

**Abstract**

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.

This project seeks to address this issue by developing a web application for news article sorting using machine learning algorithms. The primary goal is to identify and classify news articles.

**1. Introduction**

**1.1 Purpose of this High-Level Design Document**

The purpose of this High-Level Design (HLD) Document is to provide detailed insights into the project to support the coding phase. It aims to identify contradictions before coding begins and serves as a reference manual for understanding the high-level interactions of different modules.

The HLD will:

* Detail all design aspects comprehensively.
* Describe the user interface.
* Specify hardware and software interfaces.
* Outline performance requirements.
* Include design features and project architecture.

**1.2 Scope**

The HLD documentation outlines the system's structure, including database architecture, application architecture (layers), application flow (navigation), and technology architecture. The document uses non-technical to mildly technical language to ensure clarity for system administrators.

**1.3 Definitions**

|  |  |
| --- | --- |
| **Term** | **Description** |
| Database | Repository of all information monitored by the system |
| IDE | Integrated Development Environment |
| Render | Render.com, used for deployment |

**2. General Description**

**2.1 Product Perspective**

This project focuses on news article sorting using classification-based machine learning algorithms.

**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.

**2.3 Proposed Solution**

The proposed solution involves a classification-based machine learning model using algorithms such as Random Forest, Decision Tree classifier, Multinomial Logistic Regression. After evaluating various algorithms, Random Forest was found to be the most effective for our dataset. The process includes data preprocessing, such as removing stopwords, removing special characters, Lemmatization, mapping text to numerical format, creating , fitting and mapping different ml models followed by model evaluation.

**2.4 Technical Requirements**

This document outlines the requirements for sorting news articles using various technologies. The project requirements include:

* Exposing the model through an API or user interface for testing.
* Deploying the model on render.com.

**2.5 Data Requirements**

The data requirements are tailored to the problem statement:

* Preprocessing, modeling, and testing of data.
* Utilizing the BBC News dataset from Kaggle for training and testing.
* Collecting user input, news article text.

**2.6 Tools Used**

* Python programming language and libraries like NumPy, Pandas, nltk, Scikit-learn.
* VSCode as the Integrated Development Environment (IDE).
* render.com for deploying the model.
* HTML/CSS for front-end development.
* Flask for back-end development and API creation.
* GitHub

**2.7 Constraints**

The system should be user-friendly, error-free, and should not require users to understand the backend workings.

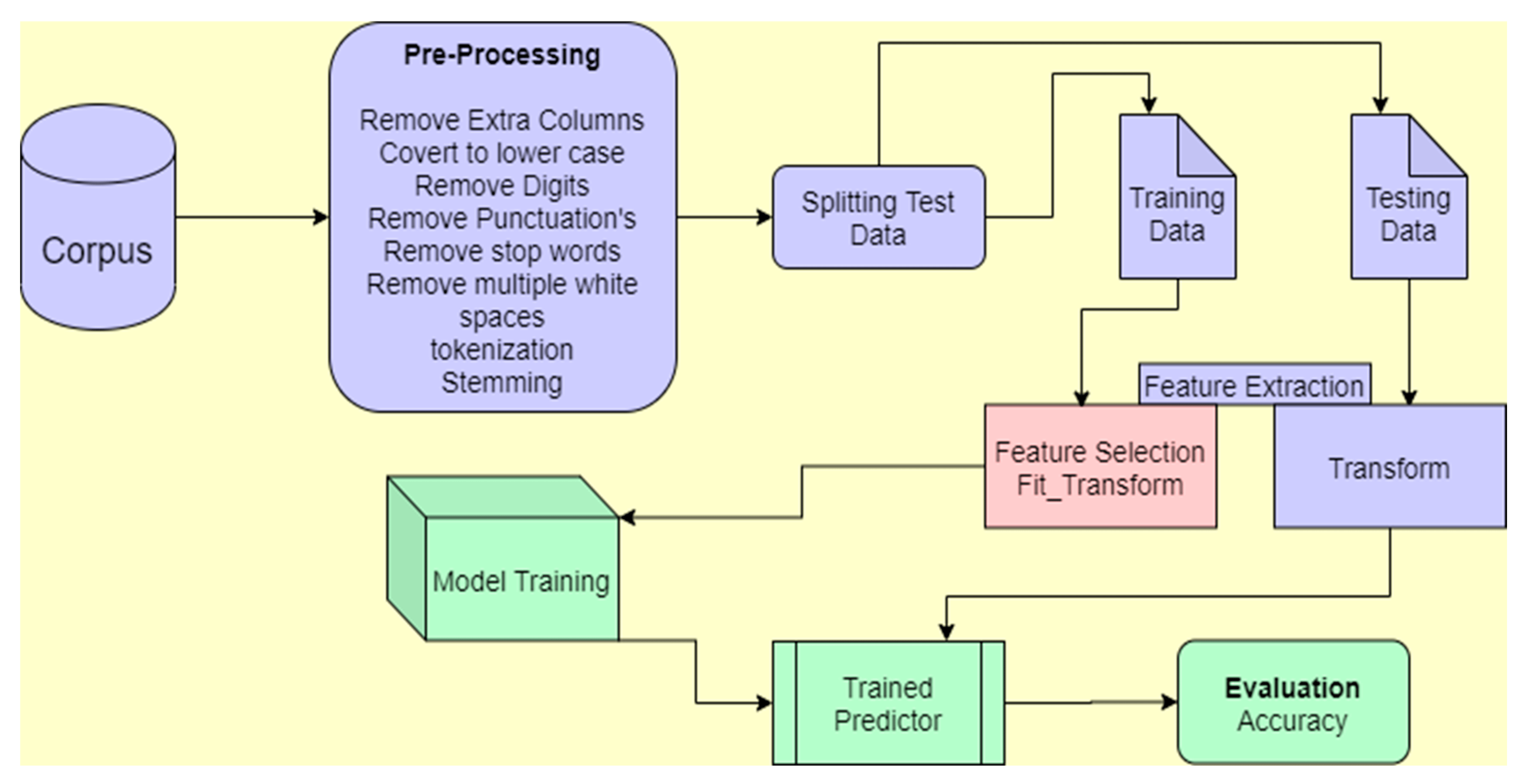
**2.8 Assumptions**

The project assumes the ability to handle new datasets for news article classification using machine learning models to identify categories.

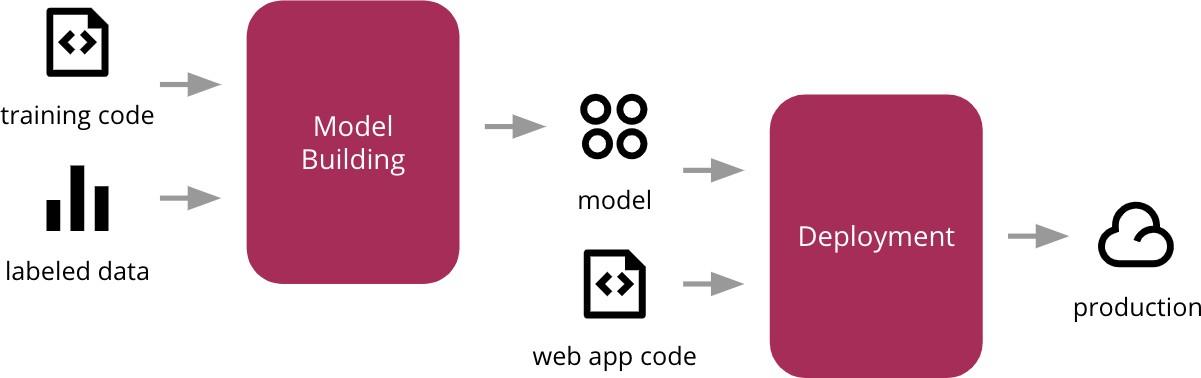
**3. Design Details**

**3.1 Process Flow**

The process flow for news article sorting using a machine learning model is as follows:



##### Model Training and Evaluation



1. **Error Handling**:
   * Display explanations for encountered errors.
   * Define errors as any deviations from normal usage.

**3.2 Performance**

The model should accurately predict news article category. The development process includes:

1. Cleaning the dataset by removing null and duplicate values.
2. Preprocessing article text data feature.
3. Creating, fitting different ml model with classification algorithms.
4. Evaluating best performing model via accuracy, precision, recall, F1-score.
5. Saving the model in pickle file format for deployment.
6. Deploying the model on render.

**3.3** Deployment.

• Render.com :

* Utilizing render for deploying the machine learning model.

**3.4 Future Enhancements**

Future enhancements to the system may include:

1. **Model Improvements**:
   * Continuously updating the model with new data to improve accuracy and adapt to emerging different news categories.
2. **Advanced Analytics**:
   * Incorporating advanced analytics and machine learning techniques, such as deep learning, neural network to enhance news category prediction.
3. **User Feedback Loop**:
   * Implementing a feedback loop where users can report false positives/negatives to improve the model.

**4. Conclusions**

The ‘News Article Sorting Project’ aims to develop a robust, scalable, and accurate system for news article sorting. By leveraging machine learning algorithms and deploying on render, the system is designed to handle large volumes of data and provide real-time news article classification. The comprehensive design details ensure that the system is user-friendly, secure, and reliable, addressing the critical need for effective news article classification in the digital age.