Low Level Design

News Article Sorting

Written by **SACHIN KUMAR G**

Contents

1. Introduction.....2

1.1 What is New Article Sorting.....2

1.2 Scope.....2

1.3 Architechture.....2

2. Project Description.....2

2.1 Data Collection.....2

2.1.2 Data Set.....3

2.2 Exploratory Data Analysis (EDA).....3

2.3 Feature Engineering.....3

2.4 Model Selection.....3

2.5 Hyperparameter Tuning.....3

2.6 Model Training.....3

2.7 Model Evaluation.....4

2.8 Model Deployment.....4

1. Introduction

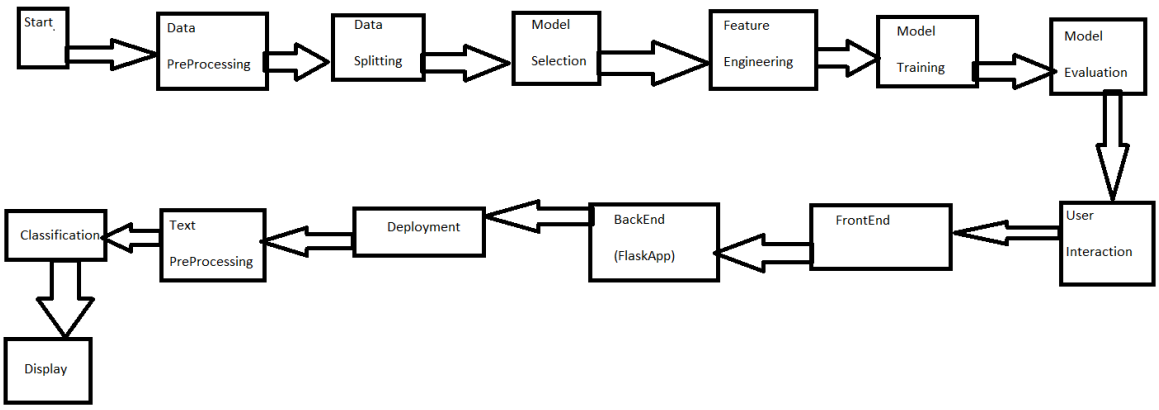
1.1 What is News Article Sorting

News article sorting is crucial for enabling efficient access to information, personalizing content delivery, organizing content repositories, identifying trends, supporting data-driven decisions, and improving overall user experiences in the digital age.

1.2 Scope:

The scope of this project involves developing a machine learning-based system to automatically categorize news articles into relevant topics, enhancing user access to tailored content and enabling potential personalized recommendations.

1.3 Architecture:



2. Project Description

2.1 Data Collection:

2.1.2 Data Set:

Download the dataset for this project from this link (https://www.kaggle.com/c/learn-ai-bbc/data)

Gather a diverse dataset of news articles with labeled categories (Finance, Sports, etc.).

2.2 Exploratory Data Analysis (EDA):

1. Perform basic analysis on the dataset to understand the distribution of categories.

2. Visualize the data to gain insights into the distribution and relationships between categories.

2.3 Feature Engineering:

1. Extract relevant features from the pre-processed text data that could help improve classification accuracy.

2. Consider using techniques like n-grams, topic modelling, and sentiment analysis as additional features.

2.4 Model Selection:

Choose appropriate machine learning models for text classification. Common choices include:

- Decision Tree
- Random Forest
- Naïve Baye's
- Non-negative matrix factorization (NMF)
- Tf-Idf CountVectorizer

2.5 Hyperparameter Tuning:

- GridSearchCV

2.6 Model Training:

1. Split your dataset into training, validation, and test sets.

2. Train your chosen model(s) using the training data and fine-tune hyperparameters.

2.7 Model Evaluation:

1. Evaluate the trained model(s) on the test set using appropriate evaluation metrics (accuracy, precision, recall, F1-score, etc.).

2.8 Model Deployment:

1. Once satisfied with the performance, deploy the trained model to a production environment.

2. Set up an API or web interface that allows users to input news articles for classification.