**DETAILED PROJECT REPORT (DPR)**

**News Article Sorting**

Index

1 . Introduction

2. Architecture

3. Workflow

3.1 Data Ingestion and Storage

3.2 Data Preprocessing

3.3 Data Transformation

3.4 Model Building

3.5 Prediction

**1 . Introduction**

**1.1 Objective:**

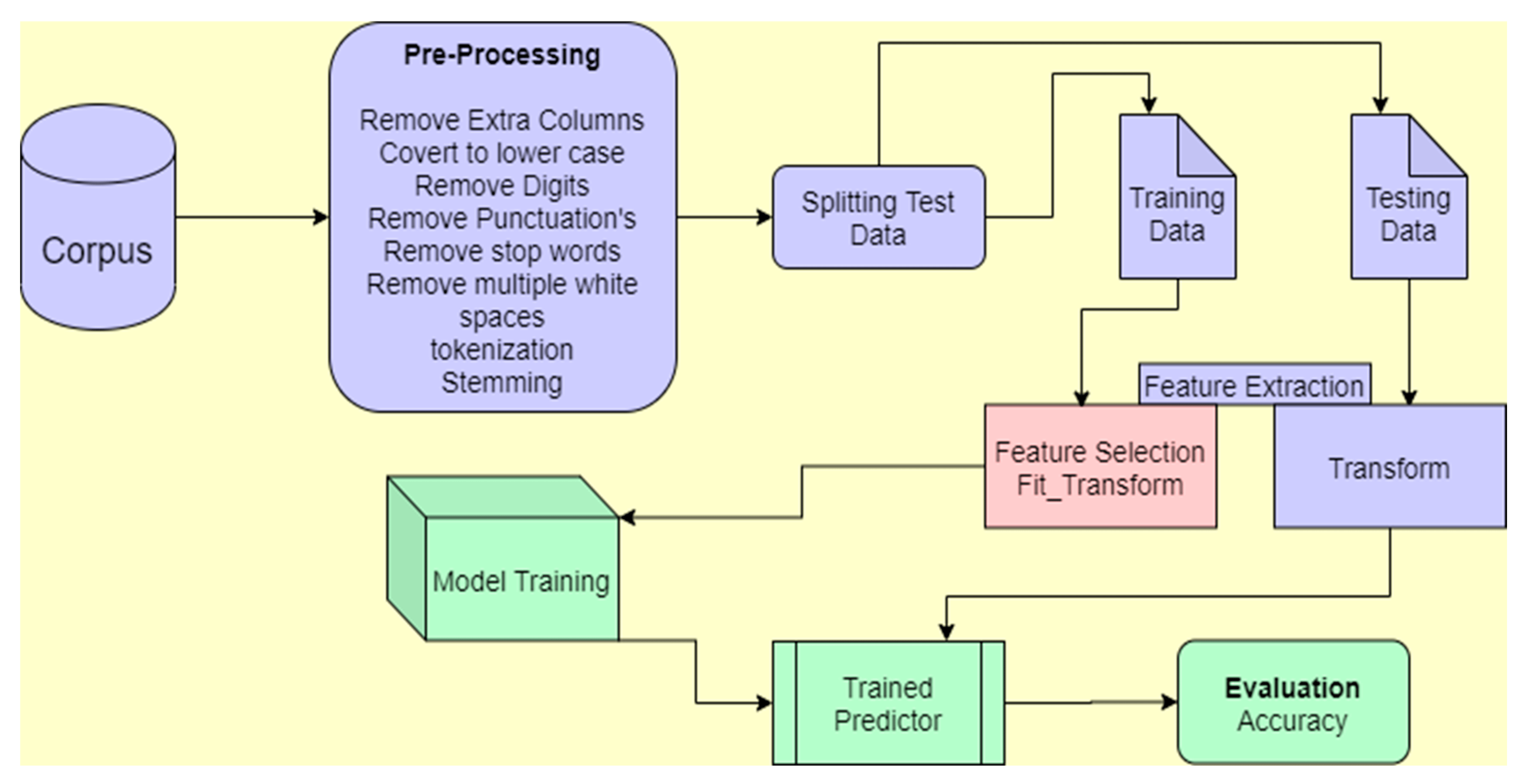
Development of a classification model for classification of news article. The model will classify news article into different categories.

**1.2 Abstract:**

IIn 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 into various categories.

**2. Architecture**



**3. Workflow**

**3.1 Data Ingestion and Storage**

1.1 Data Sources**:**

Data may come from various sources like databases, e-commerce platforms, or external sources. Ingest data from these sources. This is project of iNeuron so they provided the data for this project.

1.2 Data Storage:

We can store the data in a database or data warehouse. We can consider using technologies like PostgreSQL, MySQL, or cloud-based solutions like Amazon Redshift or Google BigQuery for scalability.

**3.2 Data Preprocessing:**

Implementing ETL (Extract, Transform, Load) processes to clean and preprocess the text data. This

includes handling missing values, removing stopwords, lemmatization.

* Text Preprocessing : Steps in preprocessing include removing stopwords, removing special characters, lemmatization, etc.
* Exploratory Data Analysis (EDA):

It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables. Tools (python libraries) like Pandas, Matplotlib, or Seaborn are used to perform EDA.

**3.3 Data Transformation :**

Explaining the data transformation phase involves detailing how raw data is processed and modified to make it suitable for consumption by downstream systems or applications.

**3.4 Model Building :**

* Model Training : Explaining the model building phase involves detailing how machine learning models are developed and trained to achieve specific objectives. In this project for news article classification, multinomial logistic regression and decision tree classifier, random forest algorithm is used for classification model.
* Model Validation : The model validation stage is a critical step in the machine learning lifecycle, and ML model is validated in this stage. This stage is designed to assess the performance and generalizability of a machine learning model. Model’s performance is evaluated by accuracy, precision, recall, F1-score.

**3.5 Prediction :**

The preprocessed and transformed features are input into the trained random forest model. The model applies its learned patterns and algorithms to generate a prediction.