



# **News Articles Sorting**

Raushan Kumar

2021MSBDA033

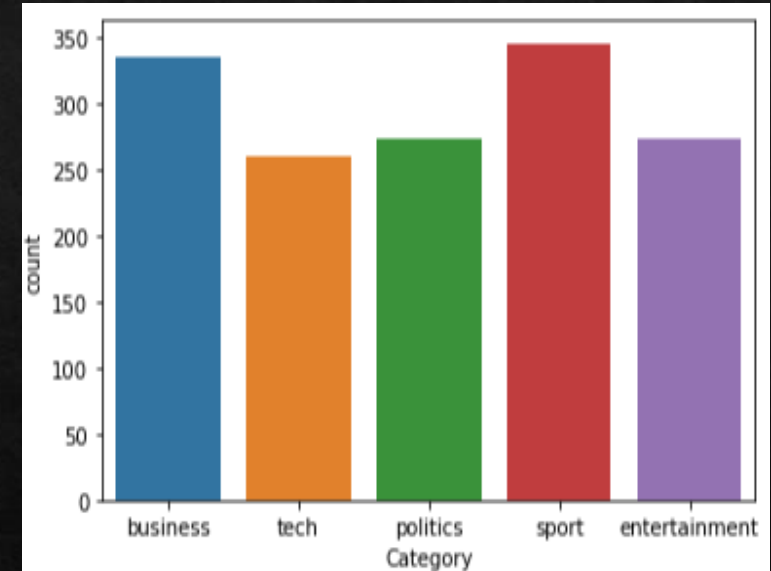
# Objective

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.

Now a days 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.

# Data Description

- The data is downloaded from kaggle i.e BBC news Dataset
- The dataset provided to us contains many rows, and 2 independent features. We aim to predict category of a news. So this clearly is a classification problem, and we will train the classification models to predict the desired outputs based on input News.
- **Article Id** – Article id unique given to the record
- **Article** – Text of the header and article
- **Category** – Category of the article (tech, business, sport, entertainment, politics)





# Data Analysis step



## DATA COLLECTION

In step 1, we collect data which is generally present in a database or on internet.



## DATA PREPROCESSING

In step 2, we preprocess the data which involves data cleaning by handling outliers, null values etc.



## EXPLORATORY DATA ANALYSIS

In step 3, we explore the data by performing univariate and bivariate analysis on the features.



## FEATURE SELECTION

In step 4, we use feature selection techniques to filter out the most important features to perform model creation



## MODEL CREATION AND EVALUATION

In step 5, we finally build models on our dataset and choose the model which gives the best accuracy.

# Data Preprocessing

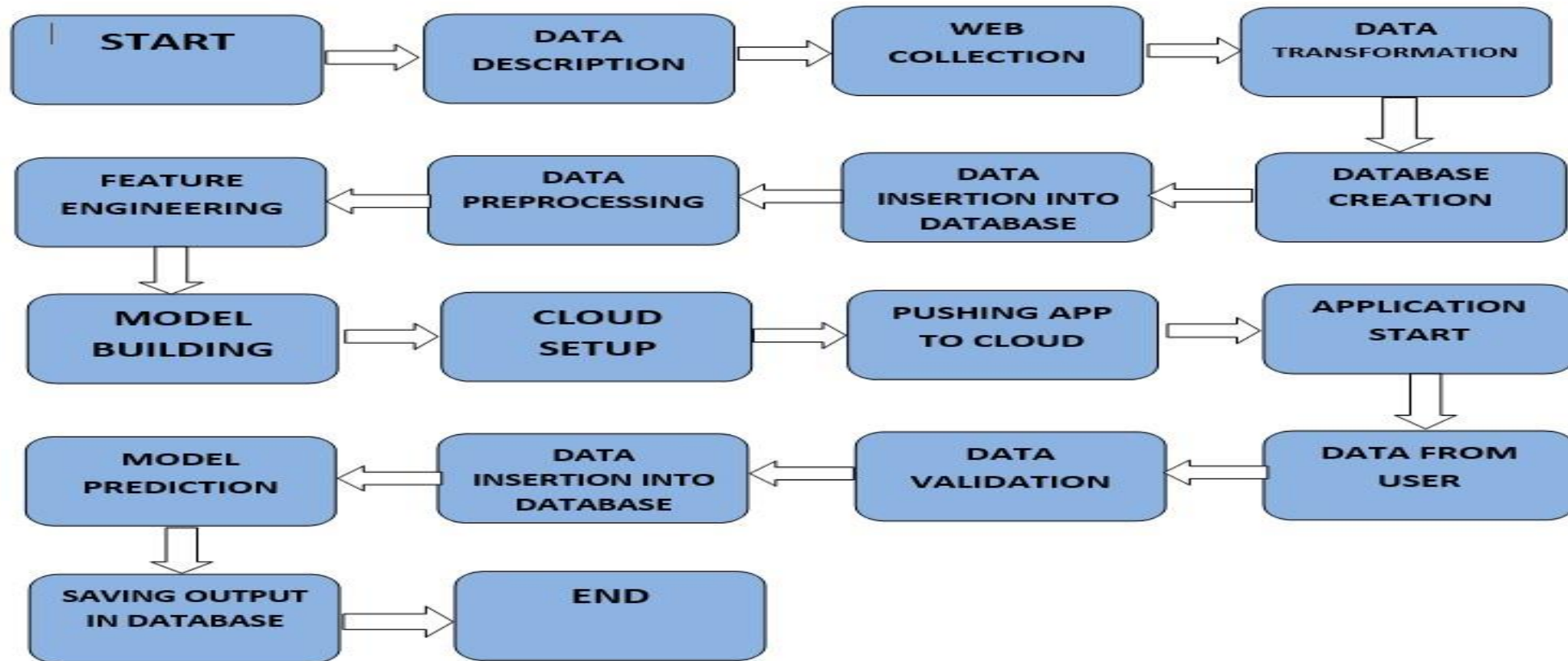
- Our data is not preprocessed and it contains a lot of punctuations and numbers. To convert raw data as a preprocessed format(like removing punctuations, numbers, single character, multiple spaces). We had created a function “preprocess\_text”
- After the raw text is preprocessing, we have to encode our labels to numerical encoding as they were categorically encoded before.

# Feature Engineering

- ◇ Checking for imbalanced dataset
- ◇ Tokenize Data
- ◇ Stemming
- ◇ Removing Stopwords
- ◇ TF IDF Vectorizer



# Architecture



# Database

- ◇ Create Database
- ◇ Create Collection
- ◇ Insertion of Data

```
def DatabaseConn(News):  
    client = pymongo.MongoClient("mongodb://localhost:27017/")  
    dataBase = client["NewsArticle"]  
    collection = dataBase["News_Article"]  
  
    record = {"News" : News}  
    collection.insert_one(record)
```





# Model Selection

- ◆ In model training , used two different Machine Learning model
- ◆ Evaluate Classification Models using Logistic Regression and Random Forest Classifier.
- ◆ First we trained Logistic Regression Model with the Accuracy of 97%.
- ◆ And Second Trained the Random Forest classifier Model with the accuracy of 94%.
- ◆ Select the model with the best score for model deployment.

# Model Deployment

- ◆ The final model is deployed using Azure Cloud platform using flask framework.

## Article Category Detection

Input an article and get its predicted category in real-time!

article

England has picked up its play in the second half of the Women's World Cup final while trailing 1-0 to Spain. The improved offense has been a reprieve for England goalkeeper Mary Earps, who has been tested throughout and was forced to make another huge save early in the second half. Aitanna Bonmatti fired a strike from outside the area in the 51st minute.

Clear

output

sport

Flag

Examples

Tech companies are at the forefront of innovation.

Recent sports events have garnered much attention.

Use via API

Built with Gradio

**Thank You**