# High Level Design

News Articles Sorting

| | |
|---|---|
| Written By | Ujjwal Kar |
| Document Version | 0.23 |
| Last Revised Date | 31.08.2022 |

## Contents

## Abstract

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. This work is a machine learning model that classifies news articles into 5 categories: business, entertainment, politics, sport, or tech. A labeled public dataset from the BBC comprised of 1490 articles is used for prediction with different algorithms. Almost every algorithm gives an accuracy of more than 90%. Complement Naive Bayes having good precision (more than 95%), recall (more than 98%), and f1-score (more than 97%) for every class, gives accuracy of 98% we use Complement Naive Bayes model in test purpose.

# 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

docType is a Natural Language Processing based article classifier that classifies news articles into 5 categories: business, entertainment, politics, sport, or tech

## 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. 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.

## 2.3. Proposed Solution

A labeled public dataset from the BBC comprised of 1490 articles is used for prediction with different algorithms. Almost every algorithm gives an accuracy of more than 90%. Complement Naive Bayes having good precision (more than 95%), recall (more than 98%), and f1-score (more than 97%) for every class, gives an accuracy of 98% we use the Complement Naive Bayes model in the test purpose.

## 2.4. Further Improvement

- Currently, docType made with a traditional machine learning algorithm with TF-IDF vectors can't classify short length (less than 10 words) sentences correctly. It is not a problem as we are classifying news articles which are not so much sort. Future plan is to improve docType using deep learning so that it can classify and sort sentences into different classes if it has any use case.

- Build a module for Javascript, Java, and C++ developer

## 2.5. Data Requirement

A public dataset from the BBC comprised of 1490 articles, each labeled under one of 5 categories: business, entertainment, politics, sport or tech.

**Dataset URL :** https://www.kaggle.com/c/learn-ai-bbc/data

**Data fields :**

- ArticleId - Article id unique # given to the record
- Article - text of the header and article
- Category - category of the article (tech, business, sport, entertainment, politics

## 2.6. Tools used





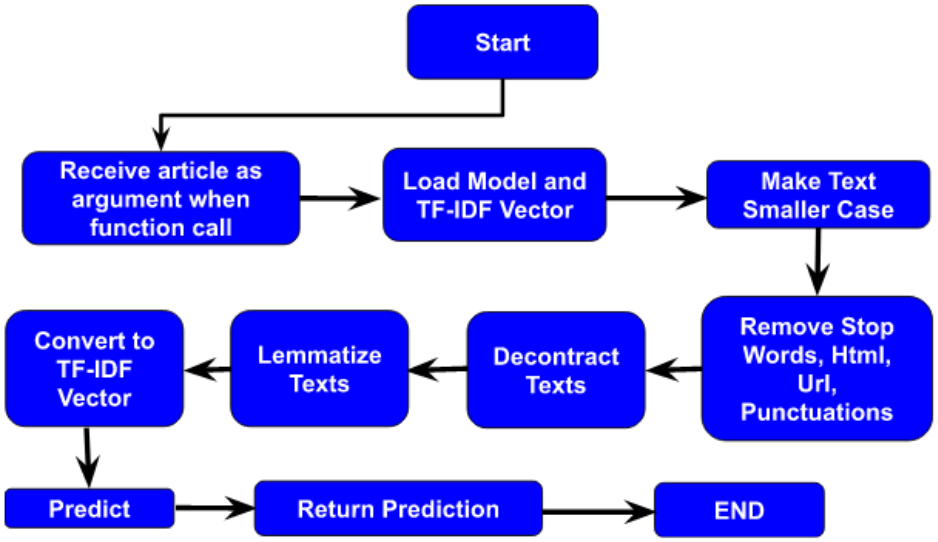

## 2.7. Constraints

docType module and API are developer-friendly which can classify articles into 5 categories. Python developers can install the docType module directly from pip, other developers can use it using REST API

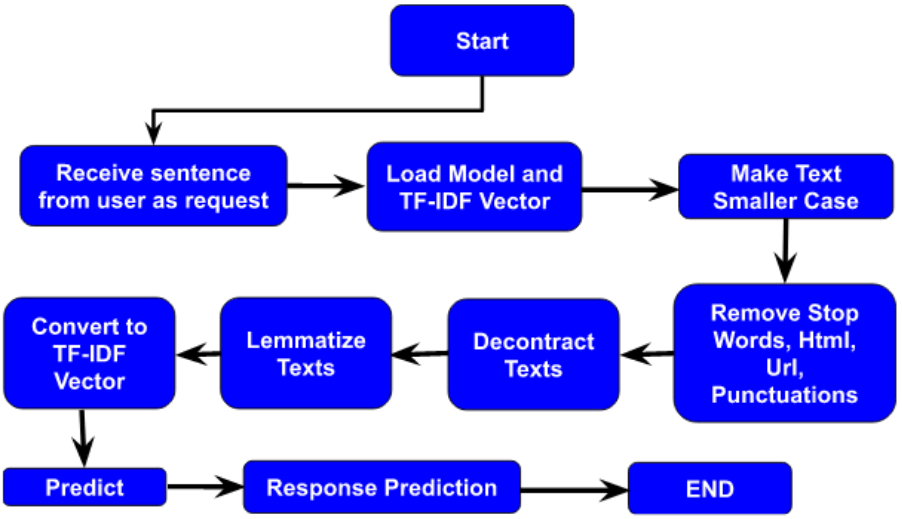
# 3. Design Details

## 3.1. Process Flow

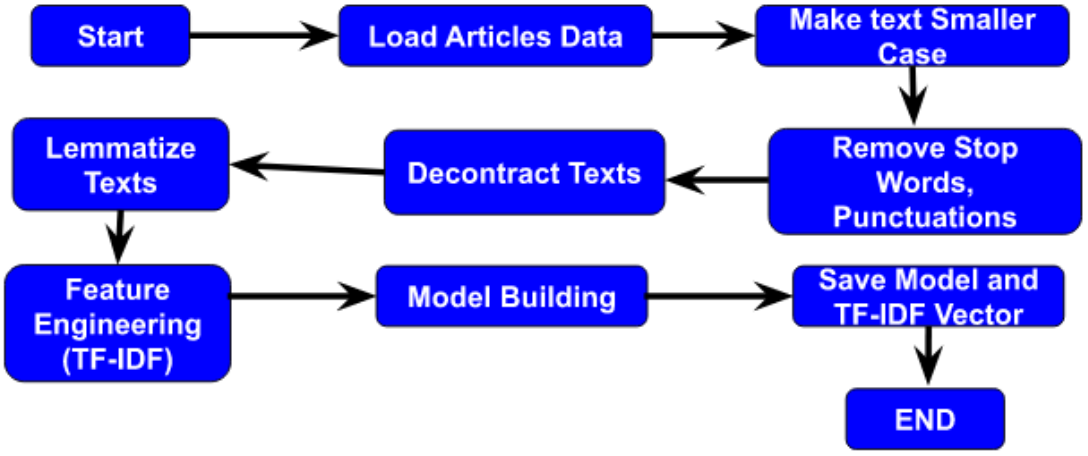
### Python Module




### Rest API

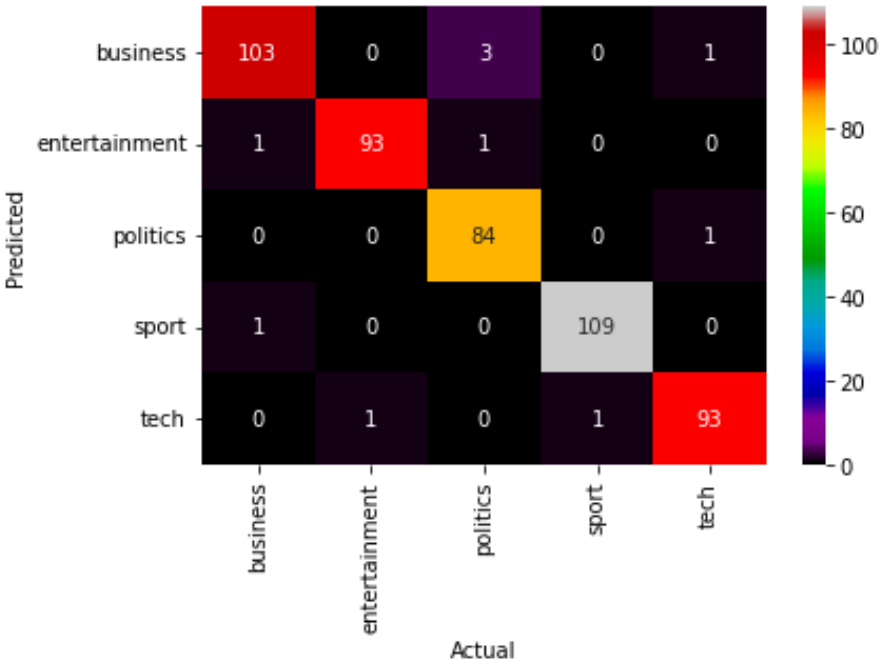


## 3.2. Training Model and Save



# 4. Performance

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| business | 0.98 | 0.96 | 0.97 | 107 |
| entertainment | 0.99 | 0.98 | 0.98 | 95 |
| politics | 0.95 | 0.99 | 0.97 | 85 |
| sport | 0.99 | 0.99 | 0.99 | 110 |
| tech | 0.98 | 0.98 | 0.98 | 95 |
| accuracy | | | 0.98 | 492 |
| macro avg | 0.98 | 0.98 | 0.98 | 492 |
| weighted avg | 0.98 | 0.98 | 0.98 | 492 |



# 5. Conclusion

The task is done with A labeled public dataset from the BBC comprised of 1490 articles used for prediction with different algorithms. docType python module allows python developers to integrate article classifiers into their python code, and REST API allows all developers to use article classifiers in their code.