NEWS ARTICLE SORTING

Revision Number: 2.0
Last date of revision: 21/11/2022

Document Version Control

Date Issued	Version	Description	Author
23/10/2022	1	Initial HLD — V1.0	Abhishek Nayak
21/11/2022	2	Updated diagrams – V 2.0	Abhishek Nayak

Contents

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.
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.

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 prior to 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.

1.3 Definitions

Term	*Description*
NLTK	Natural Language Tool Kit
TF-IDF	Term Frequency Inverse Document Frequency
IDE	Integrated Development Environment

2 General Description

2.1 Product Perspective

The ML based Automated news classification solution system helps us to identify topics of untracked news.

2.2 Problem statement

To create a ML based solution for classification of unseen news into one of the following categories :- Busiess,Tech,Politics,Sports and entertainment.

2.3 PROPOSED SOLUTION

We tested the dataset with various ML algorithms and found Naïve Bayes to be the one with the best accuracy.

2.4 FURTHER IMPROVEMENTS

The ML algorithm only provides solutions for five categories of news items. You can build a custom dataset with more categories added for better results.

2.5 Dataset

- The Dataset is a public dataset from the BBC comprised of 2225 articles, each labeled under one of 5 categories: business, entertainment, politics, sport or tech.
- The Dataset is broken into 1490 records for training and 735 for testing

2.6 Tools used

The following Frameworks were used for model building

- VSCODE is used as IDE.
- Keras is used for preprocessing the Text.
- NLTK is used for text tokenization
- For visualization of the plots, Matplotlib, Seaborn and Tableau are used.
- Front end development is done using HTML/CSS
- FastAPI is used for backend development.
- GitHub is used as the version control system.

2.7 System Requirements

- Processor – Intel corei3/AMD Reyzen 3
- RAM – 4 GB
- fastapi==0.85.0
- keras==2.10.0
- Keras-Preprocessing==1.1.2
- nltk==3.7
- numpy==1.23.3
- pydantic==1.10.2
- pyparsing==3.0.9
- regex==2022.9.13
- scikit-learn==1.1.2

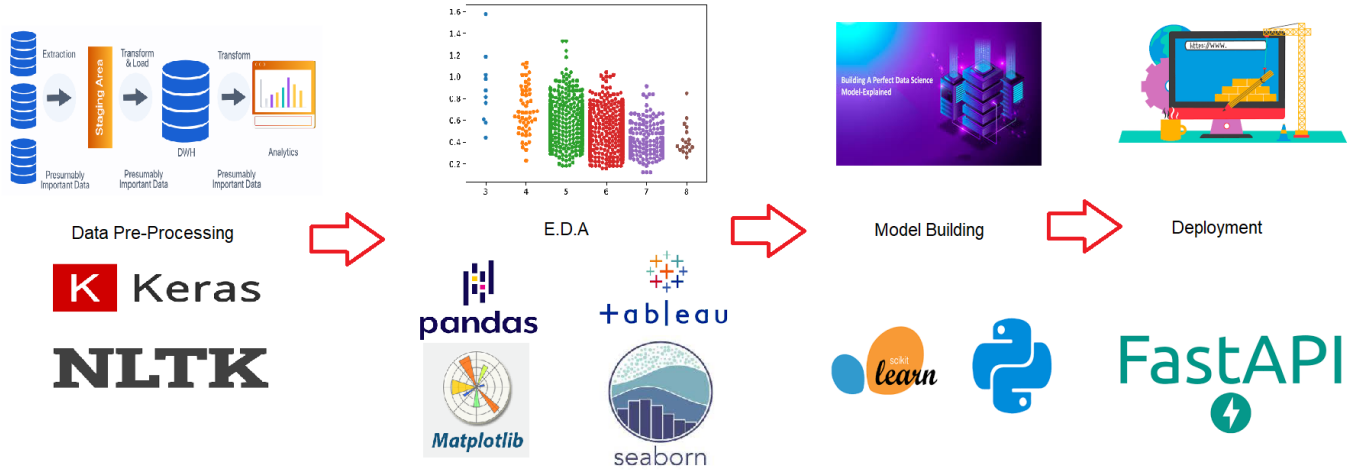
2.8 Constraints

* The model only provides solution for English News Articles.
* The model is limited only 5 categories of news items
* The text corpus should not be more than 100 words

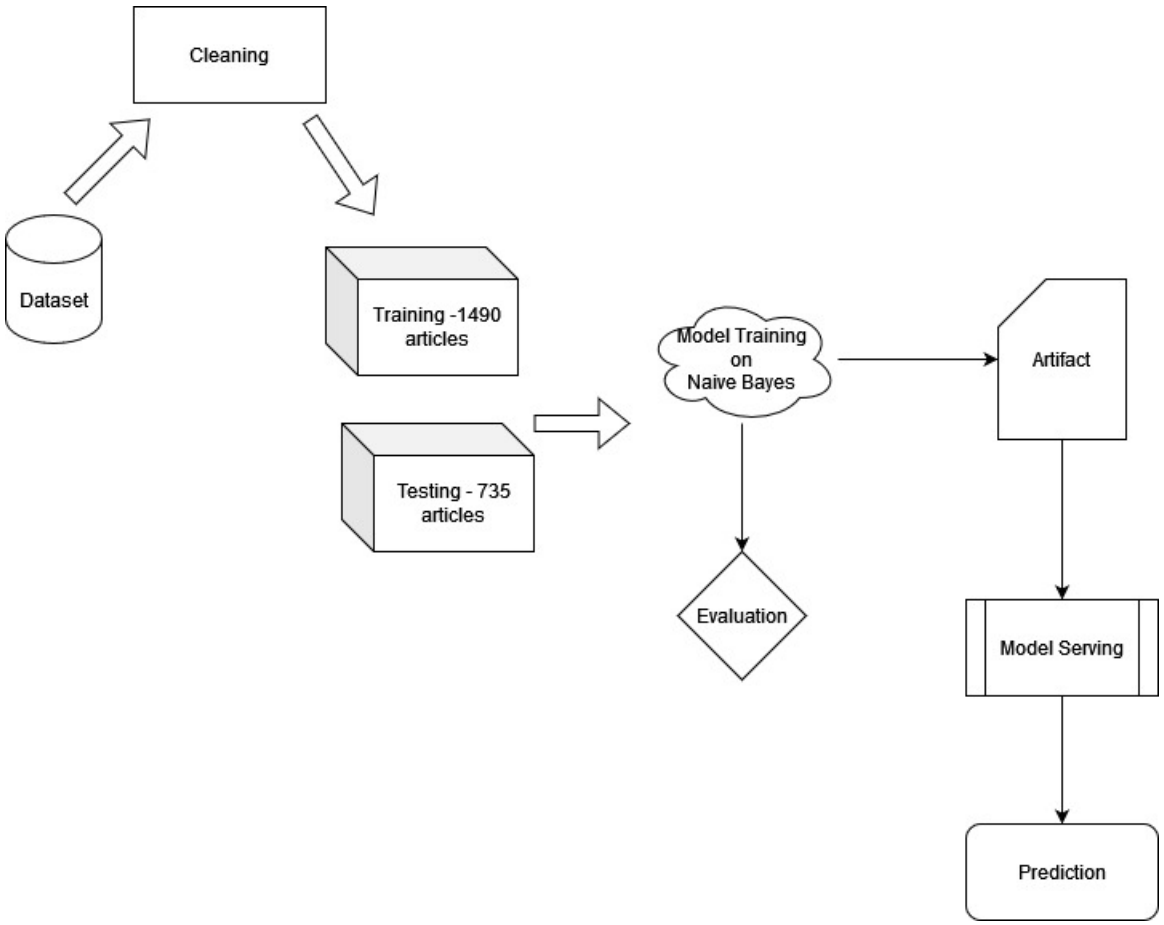
3 Design Details

3.1 Process Flow

For identifying the different types of anomalies, we will use a deep learning base model. Below is the process flow diagram is as shown below.



3.1.1 Model Training and Evaluation



4 Performance

Accuracy achieved on the Training set – 99%
Accuracy achieved in the Test Dataset – 96%

4.1 Reusability

The code written and the components used should have the ability to be reused with no problems.

4.2 Application Compatibility

The different components for this project will be using Python as an interface between them. Each component will have its own task to perform, and it is the job of the Python to ensure proper transfer of information.

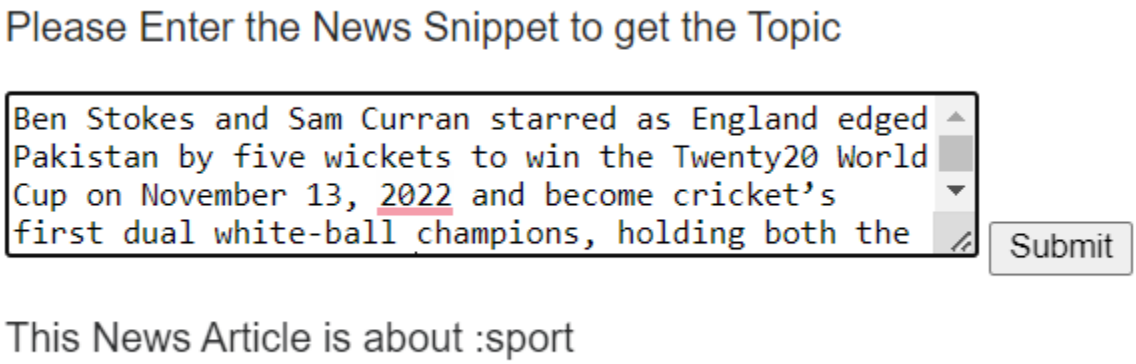
4.3 Resource Utilization

When any task is performed, it will likely use all the processing power available until that function is finished.

4.4 Deployment

5 Conclusion

We build a News article sorting algorithm using the Naïve Bayes Algorithm. The model was tested extensively and performed as per our expectations.



6 References

1. https://en.wikipedia.org/wiki/Unmanned_ground_vehicle
2. Google.com for images of UGV hardware.
3. https://www.ros.org/