High Level Design (HLD)

News Article sorting

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

1 Introduction

1.1Why this High-Level Design Document?

The purpose of this High-Level Design (HLD) Document is to add necessary details to the current project description to represent a suitable model for coding. This model 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 details.

- Describe the user interface being implemented

- Describe the hardware and software interfaces

- Describe the performance and 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, application flow (Navigations), and technology architecture. The HLD uses non-technical to mildly-technical term which should be understandable to the administrator of the system.

2 General Description

2.1 Product Perspective

In News Article sorting application, predict news category based on user's input article.

2.2 Problem Statement

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

Techniques like clustering and associating rule-based algorithms can be applied to group together similar text. The ML algorithms learn the mapping function between the text and the tags based on already categorized data. Algorithms such as SVM, Neural Networks, Random Forest are commonly used for text classification.

2.4 Technical Requirements

- Python (Programming Language version 3.7+)
- Streamlit(Python Backend Framework)
- sklearn (Machine Learning Library)
- git (Version Control Distribution)
- nltk (python library for NLP)
- pandas (Python Library for Data operations)
- NumPy (Python Library for Numerical operations)
- VS code (IDE) Azure (Cloud platform

2.5 Data Requirement

- 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)

2.6 Tools used

- Python programming language and
- frameworks such as NumPy, Pandas, Scikit-learn, Flask, Azure, Git.

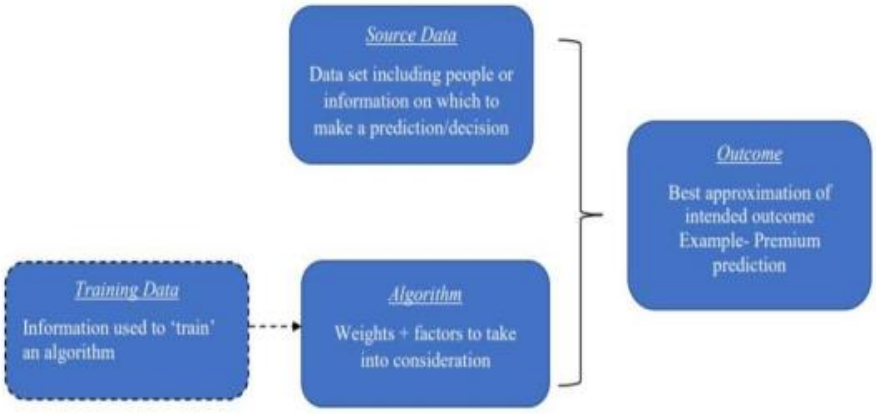
2.7 Constraints

The News article sorting must be user friendly, as automated as possible and users should not be required to know any of the workings.

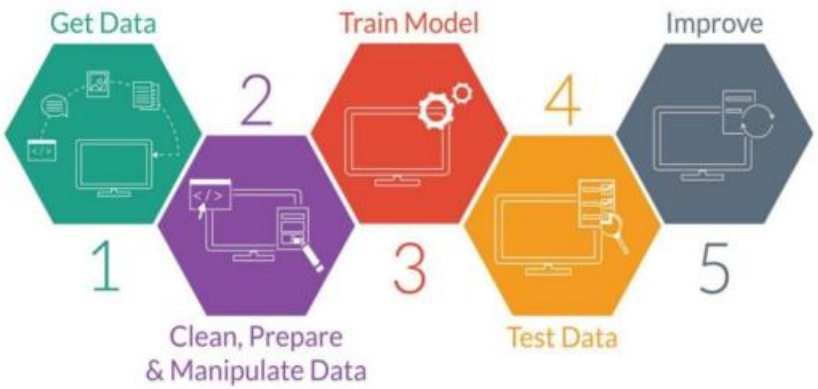
3. Design Details

3.1 Process Flow

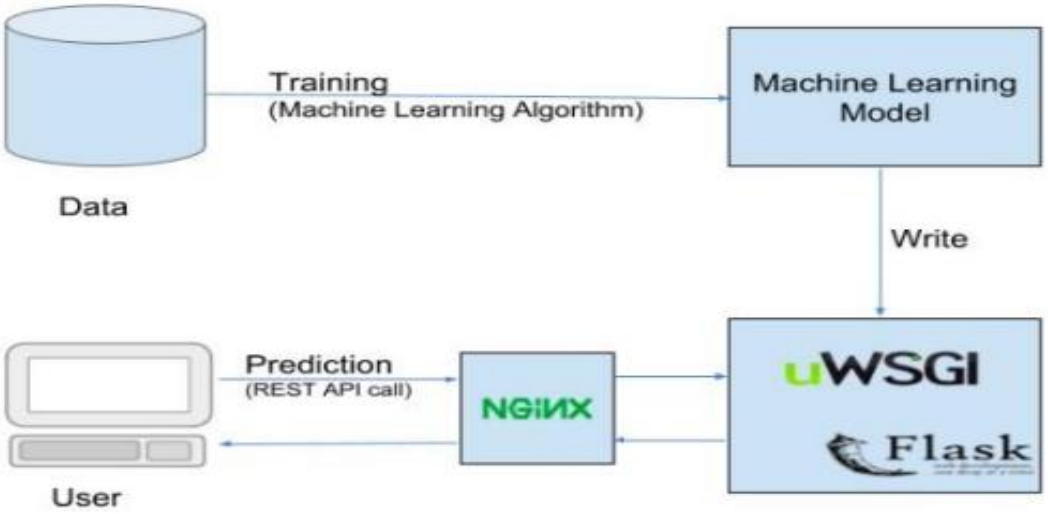
For predicting News Category, we will use Classification model. Below is the process flow diagram is as shown below.



3.1.1 Model Training and Evaluation



3.1.2 Deployment Process



3.2 Event

Log The system should log every event so that the user will know that process is running internally. Initial Step-By-Step Description:

1. The system identifies at what step logging required

2. The system should be able to log each and every system flow

3. Developer can choose logging method. You can choose database logging / File logging as well.

4. System should not hang even after using loggings. Logging just because we can easily debug issues so logging is mandatory to do.

3.3 Error Handling

Should error be encountered, an explanation will be displayed as to what went wrong? An error will be defined as anything that falls outside the normal and intended usage.

3.4 Reusability

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

3.5 Application Compatibility

The different components for this project will be using 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.

3.6 Resource Utilization

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

4 Conclusion

Background In this project, five Classification models are evaluated for individual News Article data. It has been found that MultinomialNB model which is built is the best performing model.