HIGH LEVEL DESIGN (HLD)

News Articles Sorting

REVISION NUMBER: 1.0
Last Revision Date: 09.01.2022

DOCUMENT VERSION CONTROL

Date Issued	Version	Description	Author
09/01/2023	0.0.1	Initial HLD – V1.0	Sourav Mandal

ABSTRACT:

The BBC News article classification project aims to classify news articles from the BBC News website into different categories such as politics, sports, entertainment, etc. using natural language processing techniques. In order to achieve this, a machine learning model will be developed and trained on a dataset of news articles from the website. The model will be implemented in a production environment and used to classify new news articles as they are published. The classified articles will be used for various purposes such as analysis or content curation by business stakeholders. In order to ensure the accuracy of the model, it will be evaluated using a set of test data and monitored for any necessary updates or improvements.

Objectives:

The main objective of this project is to classify news articles from the BBC News website into different categories such as politics, sports, entertainment and technology using natural language processing techniques.

Scope:

The scope of this project includes the development and implementation of the BERT model for classifying news articles from the BBC News website into different categories. It does not include the development of a front-end interface or integration with any other external systems.

Stakeholders:

The stakeholders for this project include the data scientists responsible for developing and implementing the BERT model, the data engineers responsible for collecting and preparing the data, and the business stakeholders who will be using the classified news articles for various purposes such as analysis or content curation.

Requirements:

- **Data:** The data for this project will consist of news articles from the BBC News website. The data will need to be collected and cleaned in order to be used for training the BERT model.
- **BERT model:** The BERT model will need to be trained on the cleaned data in order to accurately classify news articles into different categories.
- **Evaluation:** The model will need to be evaluated in order to determine its accuracy and identify any areas for improvement.

PROCESS FLOW:

1. Collect news articles from the BBC News website using web scraping tools.
2. Clean and prepare the collected data using Python libraries such as Pandas or NumPy.
3. Train the BERT model on the cleaned data using the Tensorflow library.
4. Evaluate the trained model using a set of test data and measure its performance using metrics such as accuracy, precision, and recall.
5. Implement the trained model in a production environment.
6. Classify new news articles as they are published on the BBC News website using the trained BERT model.
7. Monitor the performance of the model and make any necessary updates or improvements.

MODEL TRAINING AND EVALUATION:

- **Training:**
    a. The BERT model will be trained on a dataset of news articles from the BBC News website.
    b. The data will be cleaned and prepared using Python libraries such as Pandas or NumPy.
    c. The BERT model will be implemented and trained using the Tensorflow library.
- **Evaluation:**
    a. The trained model will be evaluated using a set of test data in order to determine its accuracy.
    b. Metrics such as accuracy, precision, and recall will be used to measure the model's performance.
    c. If the model's performance is not satisfactory, further development and refinement will be required.
    d. If the model performs well, it will be implemented in a production environment.
    e. Currently the trained model has an accuracy of 97% over the test set.

Solution overview:

- **Data collection and preparation:** The data engineers will collect the news articles from the BBC News website and clean the data in order to remove any irrelevant or duplicate information.
- **Model training:** The data scientists will use the cleaned data to train the BERT model using natural language processing techniques.
- **Model evaluation:** The model will be evaluated using a set of test data in order to determine its accuracy.
- **Model implementation:** The model will be implemented in a production environment and used to classify new news articles as they are published on the BBC News website.

Technology:

- **Data collection:** The data will be collected from the BBC News website using web scraping tools such as Beautiful Soup or Selenium.
- **Data preparation:** The data will be cleaned and prepared using Python libraries such as Pandas or NumPy.
- **BERT model:** The BERT model will be developed and implemented using the Transformers library.
- **Model evaluation:** The model will be evaluated using metrics such as accuracy, precision, and recall.

Resource requirements:

- **Data engineers:** The data engineers will be responsible for collecting and preparing the data for the BERT model.
- **Data scientists:** The data scientists will be responsible for developing and implementing the BERT model.
- **Computing resources:** The project will require access to a computing environment with sufficient resources to train and evaluate the BERT model.

DEPLOYMENT:

1. Set up an Azure account and create a new resource group for the project.
2. Linked the source code and the trained model from (from an S3 bucket) to the application.
3. Configure the app to use the trained BERT model for classification.
4. Test the app to ensure it is functioning correctly.
5. Deploy the app to the production environment and monitor its performance.
6. Make any necessary updates or improvements to the app as needed.

Timeline:

- **Data collection and preparation:** 2 weeks
- **Model training:** 4 weeks
- **Model evaluation:** 1 week
- **Model implementation:** 1 week

Risks:

- **Data quality:** There is a risk that the collected data may be of poor quality or contain errors, which could affect the accuracy of the BERT model.
- **Model accuracy:** There is a risk that the BERT model may not perform accurately, requiring further development and refinement.
- **Resource constraints:** There is a risk that the project may not have access to sufficient resources such as computing power or data scientists, which could impact the timeline and success of the project.