Architecture

News Article Sorting

Author: SACHIN KUMAR G

Date	Version	Description	Author
18-08-2023	1.0	Introduction, Architecture,	Sachin Kumar G

Contents:

1. Abstract:

In a digital landscape driven by the value of data, the automated classification of news articles emerges as a pivotal solution to efficiently organize vast information resources. With the proliferation of online news sources and increasing user demand for relevant content, the development of machine learning models for news article classification becomes paramount. This solution employs techniques such as clustering, rule-based algorithms, and advanced machine learning algorithms to categorize news articles into topics such as Finance and Sports. The outcome is a system capable of swiftly and accurately classifying news articles, enabling users to swiftly access pertinent information and providing the foundation for personalized content recommendations.

2. Introduction:

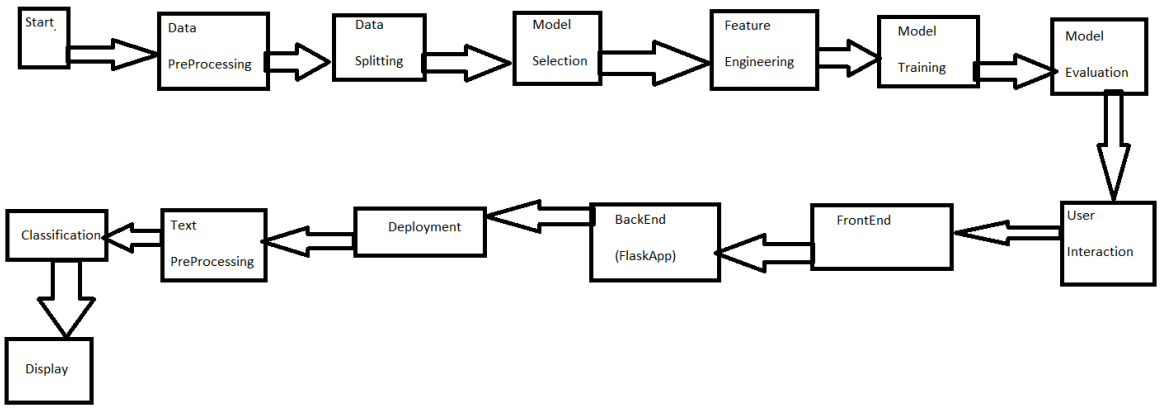
2.1 What is New Article Sorting:

News article sorting is crucial for enabling efficient access to information, personalizing content delivery, organizing content repositories, identifying trends, supporting data-driven decisions, and improving overall user experiences in the digital age.

2.2 Scope:

The scope of this project involves developing a machine learning-based system to automatically categorize news articles into relevant topics, enhancing user access to tailored content and enabling potential personalized recommendations.

3. Architecture:



4. Constraints:

We have BBC News Article Sorting Training / Testing / Submission datasets

5. Risks:

Document specific risks that have been identified or that should be considered.

6. Technical Specifications:

6.1 Dataset:

Dataset	**Finalized**	**Source**
News Articles Sorting	Yes	https://www.kaggle.com/c/learn-ai-bbc/data

6.1.1 Articles Datasets Overview:

Consists of 3 different tables. BBC News Sample Solution.csv have ArticleID and Category, BBC News Test.csv have ArticleID and Text, BBC News Train.csv have ArticleID, Text, Category

In [5]:
```
Training_Set.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1490 entries, 0 to 1489
Data columns (total 3 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   ArticleId  1490 non-null   int64
 1   Text       1490 non-null   object
 2   Category   1490 non-null   object
dtypes: int64(1), object(2)
memory usage: 35.1+ KB
```

In [6]:
```
Testing_Set.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 735 entries, 0 to 734
Data columns (total 2 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   ArticleId  735 non-null    int64
 1   Text       735 non-null    object
dtypes: int64(1), object(1)
memory usage: 11.6+ KB
```

In [7]:
```
Sample_Solution_Set.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 735 entries, 0 to 734
Data columns (total 2 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   ArticleId  735 non-null    int64
 1   Category   735 non-null    object
dtypes: int64(1), object(1)
memory usage: 11.6+ KB
```

6.1.2 BBC News Train.csv

[2]:
```
print("Training Data :- ")
Training_Set.head()
```

```
Training Data :-
```

[2]:

	ArticleId	Text	Category
0	1833	worldcom ex-boss launches defence lawyers defe...	business
1	154	german business confidence slides german busin...	business
2	1101	bbc poll indicates economic gloom citizens in ...	business
3	1976	lifestyle governs mobile choice faster bett...	tech
4	917	enron bosses in $168m payout eighteen former e...	business

6.1.3 BBC News Test.csv

[3]:
```
print("Testing Data :- ")
Testing_Set.head()
```

```
Testing Data :-
```

[3]:

	ArticleId	Text
0	1018	qpr keeper day heads for preston queens park r...
1	1319	software watching while you work software that...
2	1138	d arcy injury adds to ireland woe gordon d arc...
3	459	india s reliance family feud heats up the ongo...
4	1020	boro suffer morrison injury blow middlesbrough...

6.1.4 BBC News Sample Solution.csv

```
In [4]: print("Sample_Solution_Set :- ")
        Sample_Solution_Set.head()
```

Sample_Solution_Set :-

Out[4]:

	ArticleId	Category
0	1018	sport
1	1319	tech
2	1138	business
3	459	entertainment
4	1020	politics

```
In [5]:
```

6.2 Deployment

- Flask

7. Technology Stack:

FrontEnd	HTML/CSS
BackEnd	Python
Deployment	Flask

8. Proposed Solution:

Based on the Kaggle and Vidhya Analytics Research Articles, if we are using News Articles Sorting to classify the articles based on the category like sport, business, tech etc, then we might want to consider using Random Forest Classifier and NMF (Non – negative Matrix Factorization) Technique.

9. Model Training/Validation Workflow: