High Level Design (HLD)

News Articles Sorting

Revision Number: 2.0

Document Version Control

Version	Description	Author
1	Initial HLD - VI .0	Sachin Kumar G

Contents

1. Abstract

2. Introduction

   2.1 Why this News Article Sorting

   2.2 Scope

3. General Description

   3.1 Problem Statement

4. Proposed Solution

   4.1 Data Requirements

5. Tools Used

   5.1 Constraints

6. Assumption

7. Design Details

   7.1 Process Flow

      7.1.1 Proposed methodology

      7.1.2 Deployment Process

8. Error Handling

9. Performance

10. Reusability

11. Application Compatibility

12. Resource Utilization

13. Deployment

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 Why this News Articles Sorting?

The sorting or classification of news articles serves several essential purposes in today's information-rich environment:

1. Efficient Information Access: With the sheer volume of news articles published daily, efficient sorting allows users to quickly find and access articles relevant to their interests. This saves time and ensures that users can stay informed without being overwhelmed by the vast amount of available content.

2. Personalization: Automated sorting enables the delivery of personalized content recommendations to users based on their past reading habits and preferences. This enhances user engagement and satisfaction by providing them with content that aligns with their interests.

3. Content Organization: For news companies and publishers, sorting articles into categories or topics facilitates effective content organization. This helps in maintaining a structured content repository, making it easier for users to navigate and explore various subjects.

4. Targeted Advertising: Effective article sorting provides insights into user interests, allowing for better-targeted advertising. Advertisers can align their messages with specific categories, ensuring that ads reach the relevant audience.

5. Trend Identification: By categorizing articles, patterns and trends in news coverage can be identified more easily. This is valuable for journalists, analysts, and researchers who can quickly gather insights from the categorized data.

6. Data Mining: Sorted news articles can be used for data mining and analysis purposes. Researchers can analyze the prevalence of specific topics over time, sentiment trends, and correlations between different categories.

7. Enhanced User Experience: Offering users the ability to find articles on diverse topics through effective sorting improves their overall experience on news platforms, potentially leading to increased user retention.

8. Automated Content Generation: Sorted articles can be inputs to automated content generation systems that produce summaries, highlights, and even new articles on related subjects.

9. Real-time Insights: Real-time sorting allows users to stay up-to-date with the latest news in specific categories, fostering informed decision-making and timely reactions

10. Media Monitoring: Organizations and individuals can use sorted news articles for media monitoring purposes. They can track mentions of their brand, competitors, or specific keywords across different news sources.

In summary, 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 General Description

3.1 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.

4 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.

4.1 Data Requirements

To successfully develop and train a machine learning model for news article classification, the following types of data are required:

1. Labeled News Articles: A diverse dataset of news articles with associated labels or categories (e.g., Finance, Sports, Politics). This forms the foundation for training the model to learn the relationships between text content and corresponding topics.

2. Textual Data: The actual content of the news articles, preferably preprocessed to remove unnecessary noise (e.g., HTML tags, special characters) and converted into a suitable numerical representation (e.g., TF-IDF, word embeddings).

3. Labels or Categories: The categories or topics that each news article belongs to. These labels will be used to train the model to recognize patterns in the text that correspond to specific topics.

4. Validation and Test Sets: Separate datasets to validate the performance of the trained model and to assess its generalization on new, unseen data

5. Optional: Unlabeled Data for Pre-training: In some cases, having a larger set of unlabeled news articles can be useful for pre-training language models like BERT or GPT, which can then be fine-tuned for your specific classification task.

6. Metadata: Additional metadata about the news articles, such as publication date, author, source, etc., could be useful for improving classification accuracy and for potential analysis.

It's important to note that the quality and diversity of the dataset are crucial factors in determining the success of your machine learning model. A well-labeled, balanced dataset with representative examples from each category will help your model generalize better to new, unseen data. Additionally, ensure that the text data is accurately preprocessed to retain the relevant information while removing noise that could hinder the model's learning process.

5. Tools used

Python programming language and frameworks such as NumPy, Pandas, Scikit-learn, WordCloud, nltk and tqdm are used to build the whole model.

TensorFlow

Flask

- Visual Studio and Jupyter notebook is used as IDE.
- For visualization of the plots, Matplotlib, Seaborn and Plotly are used.
- Flask is used for deployment of the model.
- Tableau/Power BI is used for dashboard creation.
- Front end development is done using HTML/CSS
- GitHub is used as version control system.

5.1 Constraints

The News Articles Sorting or Classification must be user friendly, as automated as possible and users should not be required to know any of the working knowledge.

6 Assumption

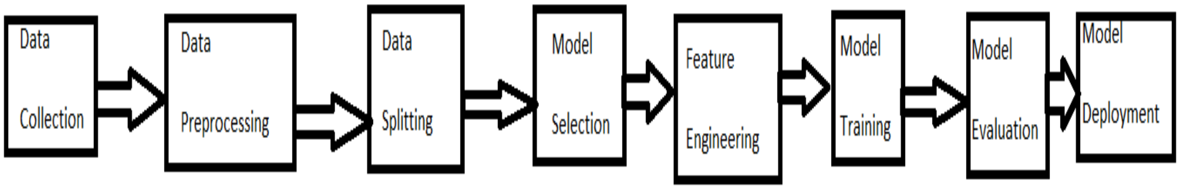
The main objective of the project is to implement Each news article belongs to one and only one predefined category.

7 Design Details

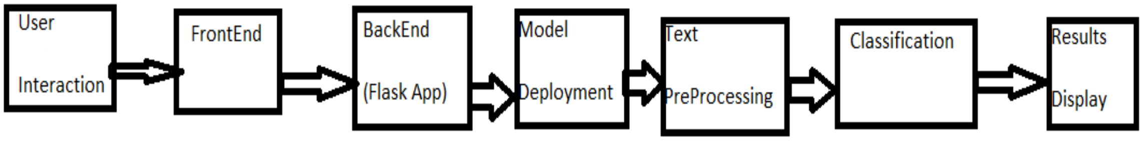
7.1 Process Flow

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.

7.1.1 Proposed methodology



7.1.2 Deployment Process



8 Error Handling

Should errors 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. And we used StackOverFlow and ChatGpt for clearing error in our coding.

9. Performance

The actual performance will depend on factors such as the quality and diversity of the training data, the choice of machine learning algorithms, feature engineering, and hyperparameter tuning. It's important to strike a balance between precision and recall based on the specific goals of the application. For instance, in news classification, you might prioritize precision if you want to ensure that articles are accurately assigned to categories, or you might prioritize recall if you want to capture as many relevant articles as possible.

10 Reusability

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

11 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.

12 Resource Utilization

When any task is performed, it will likely use all the articles will be available until that classinfication is finished.

13 Deployment

Deployment will be done by using Flask web API.