TITLE OF THE PROJECT

Genre Classification from News Title

Subject – Project Report

By Aritra Mukherjee

UID – TNU2021053100012

DEPT – CSE AI-ML

Team Members of – Group 2

1. Aritra Mukherjee

Genre Classification from News Title: Project Report

CHAPTER 1: PROJECT OVERVIEW

Objective: The project "Genre Classification from News Title" aims to develop a machine learning model that automatically categorizes news articles by their headlines. Utilizing text classification techniques and natural language processing, it identifies patterns in headlines to classify them into genres like Politics, Sports, and Business. The project investigates different feature extraction methods and machine learning algorithms to optimize accuracy. Ultimately, it provides a practical solution for efficiently organizing news content, enhancing user experience.

Research Question: "How can I efficiently categorize news headlines into relevant genres to help readers quickly find content of interest?"

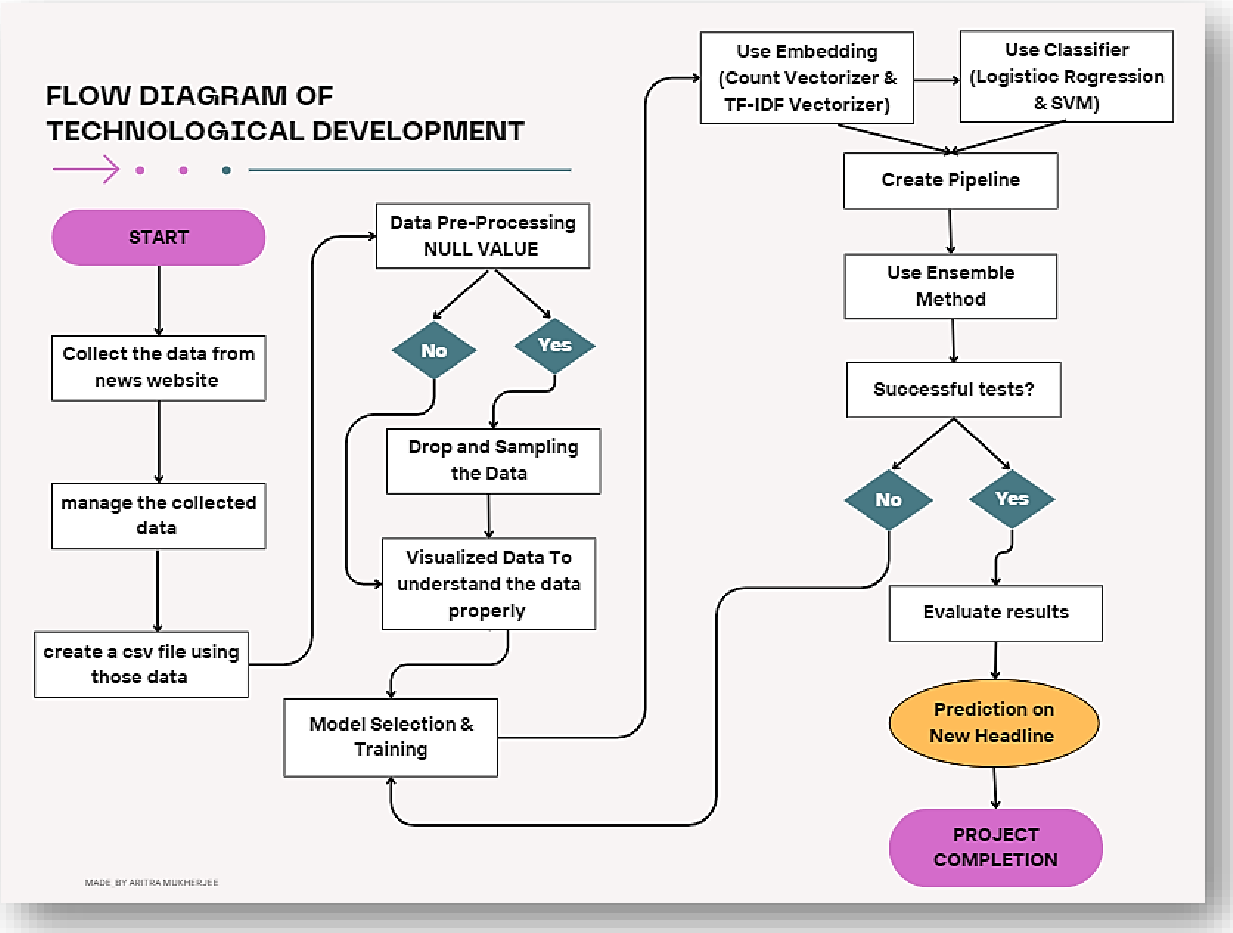
CHAPTER 2: SYSTEM ARCHITECTURE

The architecture follows a systematic approach to data collection, preprocessing, feature extraction, and genre classification. The architecture can be broken down as follows:

1. **Data Collection**
   - Fetch News headlines by Web Scraping Method using Beautiful Soup from multiple online news sources.
   - Handle data and created a large dataset, covering various genres such as Politics, Business, and Sports, Technology.
2. **Data Preprocessing**
   - The collected headlines are cleaned and tokenized to remove noise such as punctuation, stop words, and irrelevant characters.
3. **Model Training**
   - **Feature Extraction**: Using embeddings, Count Vectorizer and TF-IDF Vectorizer are applied to convert the textual data into numerical representations suitable for machine learning models.
   - **Classification Pipeline**: A machine learning pipeline is implemented, combining feature extraction with classifiers. Logistic Regression and Support Vector Machine (SVM) are used to train and predict news genres.
   - **Ensemble Method**: An ensemble approach is utilized to enhance classification performance by combining predictions from multiple classifiers for improved accuracy.

4. **Data Visualization & Evaluation**:

- ○ Visualizations are generated to analysed the distribution of genres and model performance, providing insights into the classification results & Model performance is evaluated using accuracy to ensure reliable genre predictions.


FLOW DIAGRAM OF TECHNOLOGICAL DEVELOPMENT
START
Collect the data from news website
manage the collected data
create a csv file using those data
Data Pre-Processing NULL VALUE
No
Yes
Drop and Sampling the Data
Visualized Data To understand the data properly
Model Selection & Training
Use Embedding (Count Vectorizer & TF-IDF Vectorizer)
Use Classifier (Logistioc Rogression & SVM)
Create Pipeline
Use Ensemble Method
Successful tests?
No
Yes
Evaluate results
Prediction on New Headline
PROJECT COMPLETION
MADE BY ARITRA MUKHERJEE

CHAPTER 3: DATA COLLECTION

- **Source Selection**: News articles are collected from two prominent Indian news websites: The Times of India and Hindustan Times, ensuring a diverse range of genres.

- **Web Scraping Tool**: Beautiful Soup, a Python library, is employed for web scraping to extract relevant data from the HTML structure of the web pages.

- **Target Data**: The focus is on collecting headlines along with their respective genres, capturing the essence of the news content.

- **Scraping Process**:

  - ○ **HTTP Requests**: Send HTTP requests to the news websites to retrieve HTML content.

- o **HTML Parsing**: Utilize Beautiful Soup to parse the HTML and locate the specific tags containing the headlines and genres.

• **Data Storage**: The extracted headlines and genres are stored in a structured format, save as a News_Headlines CSV file, for further processing and analysis.

• **Data Quality**: Efforts are made to ensure the quality of the data by filtering out duplicates and irrelevant articles during the scraping process.

- **Challenges**
  1. **Data Diversity and Bias**: Ensuring a balanced representation of different genres can be challenging, as certain topics may dominate the news landscape, leading to biased training data that affects model performance.
  2. **Dynamic Web Structure**: The websites' HTML structures may change over time, which can break the scraping script and require frequent updates to the web scraping code to maintain consistent data collection.

CHAPTER 4: DATA PREPROCESSING

1. **Check Genre Categories**: The first step involves examining the dataset to identify the different genre categories present in the news headlines, ensuring that they align with the project’s objectives.
2. **Identify Null Values**: A thorough check for null or missing values in the dataset is performed, which could negatively impact the training process.
3. **Remove Null Values**: Any rows containing null values are dropped from the dataset to maintain data integrity and ensure complete entries for analysis.
4. **Data Cleaning**: Textual data is cleaned to remove any irrelevant characters, punctuation, and special symbols that may interfere with the classification process.
5. **Balance the Dataset**: The dataset is analysed for class distribution. If any genre category is underrepresented, techniques such as oversampling or under sampling are employed to achieve a balanced dataset, enhancing the model's ability to generalize across all genres.
6. **Final Review**: A final review of the pre-processed dataset is conducted to confirm that it meets the quality standards required for subsequent feature extraction and model training.

CHAPTER 5: MODEL TRAINING

• **Model Selection**: Logistic Regression and Support Vector Machine (SVM) are used due to their effectiveness in handling text classification tasks, with Logistic Regression providing probabilistic outputs and SVM offering robustness in high-dimensional feature spaces.

• **Feature Extraction Process**:

- **Count Vectorizer**: Transforms the text into a matrix of token counts, capturing the frequency of each word in the headlines.
- **TF-IDF Vectorizer**: Weighs the importance of words based on their occurrence in the dataset, reducing the influence of common terms while highlighting informative ones.

- **Classification Pipeline**:

  - **Logistic Regression**: Utilized for its effectiveness in classifying text data by predicting the probability of each headline belonging to a specific genre.
  - **Support Vector Machine (SVM)**: Implemented to identify the optimal hyperplane for separating different classes in the feature space.

- **Ensemble Method**:

An ensemble technique is applied to improve overall classification performance.

  - **Voting Classifier**: This method combines predictions from Logistic Regression and SVM, with the final genre determined by majority voting, enhancing the reliability of the predictions.

CHAPTER 6: DATA VISUALIZATION AND EVALUATION

- **Visual Representation:**

The aggregated data will be visualized using:

  - **Pie Charts**: To represent the distribution of news genres within the dataset. This visualization provides a clear view of the proportion of each genre, aiding in understanding the dataset's composition.

- **Model Evaluation**:

  - **Accuracy Score**: The model's performance is assessed using the accuracy score, indicating the proportion of correctly classified headlines among the total predictions. This metric provides a straightforward evaluation of the model's overall effectiveness in genre classification.

CHAPTER 7: CONCLUSION

- **Summary:**

This project successfully developed a genre classification system for news titles using advanced machine learning models. By effectively categorizing headlines into genres like Politics, Business, and Sports, the model provides valuable insights into news content. The integration of Logistic Regression and Support Vector Machine (SVM) enhances classification accuracy. Overall, this system demonstrates the potential of machine learning to streamline news organization for readers.

- **Future Work:**

The project can be further enhanced by:

- **Improved Feature Extraction**: Exploring advanced natural language processing techniques, such as word embeddings and transformer-based models, to enhance feature extraction and improve classification accuracy.
- **Model Optimization**: Fine-tuning hyperparameters and experimenting with different ensemble methods to achieve better performance and reliability in genre classification.
- **User Feedback Integration**: Implementing mechanisms to gather user feedback on classification results, allowing for continuous improvement of the model based on real-world use.
- **Deployment for Public Use**: Developing a user-friendly interface or application to make the genre classification system accessible to a wider audience, enabling easy interaction with the model.