News Categorization

Data Collection Data Exploration

- Source a dataset containing a diverse collection of news articles labeled as ‘business’, ‘sports’, ’entertainment’, ’tech’ and ‘politics’. Trusted platforms like Kaggle or datasets provided by research institutions can be valuable resources.
- Explore the dataset to understand its structure, size, and characteristics. Identify potential challenges and patterns.

Preprocessing

- Handle missing values.
- Clean text data (remove irrelevant characters, punctuation, etc.).
- Tokenization: Break down the text into individual words or tokens.
- Lowercasing, stopword removal, stemming, or lemmatization (text normalization steps).
- Handling numerical values, removing URLs, special characters,Balancing Classes,Encoding Labels and dealing with rare words.
- Convert text into a suitable format for analysis.

Exploratory Data Analysis

- After preprocessing, conduct EDA to gain insights into the dataset.
- Visualize the data to identify trends and patterns.
- Generate word clouds, histograms, or other relevant visualizations to understand the distribution of words or features.

Embedding (Feature Engineering)

- Once the text is preprocessed, you can use embedding techniques to represent words or phrases as vectors. Embeddings capture semantic relationships between words and are often used to convert text data into a format suitable for machine learning models.
- Techniques like Word Embeddings (Word2Vec, GloVe), TF-IDF vectorization,Embeddings from Language Models (BERT, GPT), or custom embeddings trained on your specific dataset can be applied in this phase.

Model Selection

- Choose a machine learning model suitable for text classification tasks. Common models include logistic regression, Naive Bayes, Support Vector Machines (SVM),Gradient Boosting Models or use more advanced models like deep neural networks with Long Short-Term Memory (LSTM) Networks or Transformer Models or LLM models.

Model Training

- Train the selected model using the preprocessed and embedded data.
- Splitting the Dataset : Divide your dataset into training and testing sets
- Initialization and Configuration : Initialize the selected model and configure its parameters
- Training the Model : Fit the model to the training data.

Model Evaluation

- Model evaluation involves assessing how well your trained model performs on unseen data. The choice of evaluation metrics depends on the nature of your classification problem.
- Confusion Matrix
- Accuracy
- Precision, Recall, and F1-Score
- Classification Report
- Cross-Validation

Model Fine-Tuning

- Model fine-tuning involves adjusting the hyperparameters or exploring different architectures to improve the performance of your machine learning model.
- Grid Search
- Randomized Search
- Explore Different Neural Network Architectures
- Use Pre-trained Models and Fine tune it with your own data

News Research Tool

- A user-friendly News Navigator Tool designed for classification and effortless information retrieval.
- Users can input the article , the LLM model will predict the article belongs to which category (business,sports,entertainment,tech,politics) and it will give a summary of that article also.
- Process article content through the fine tuned classification LLM model.
- Interact with the text generation language model (LLM) from Palm, and it will provide a summary of the article given by the user.

Deployment

- Once satisfied with the model's performance, deploy it for real-world use. Here we can use the python libraries like Streamlit or Chainlit.
- Deploying a machine learning model for real-world use involves making it accessible to users or systems.

Monitoring and Maintenance

- Continuously monitor the model's performance in a production environment. Update the model as needed to maintain its effectiveness over time.

Reports and Insights

- You can generate various reports and insights to provide a comprehensive understanding of the classified articles.
- Visualize the distribution of news categories over time.
- Geographical Spread Analysis
- Common Keywords and Phrases