**Indian English News (2023) Analysis and Classification**

**Introduction:**

The project aims to analyse and classify Indian English news articles from a dataset collected in 2023. The project leverages machine learning techniques to categorize news articles into different categories based on their headlines.

**Dataset:**

The dataset used in this project consists of English news articles collected in 2023. Each news article in the dataset is labelled with a category based on its content. The dataset includes features such as Headline and News Categories.

**Technologies Used:**

- Python: The primary programming language used for data pre-processing, analysis, and modelling.

- Libraries:

- Pandas: For data manipulation and analysis.

- NumPy: For numerical computing.

- Matplotlib and Seaborn: For data visualization.

- NLTK: For natural language processing tasks such as text pre-processing.

- Scikit-learn: For machine learning tasks including data pre-processing, model training, and evaluation.

- Keras: For building and training deep learning models.

**Workflow:**

1. Data Preprocessing:

- Loading the dataset from CSV files.

- Cleaning and preprocessing text data by removing special characters, symbols, and stopwords.

- Tokenization: Splitting the text into individual words.

- Creating word sequences and calculating the number of words in each headline.

- Handling articles with insufficient text data.

2. Exploratory Data Analysis (EDA:

- Visualizing the distribution of news categories.

- Analyzing the length of headlines for different news categories.

3. Model Development:

- Splitting the dataset into training and testing sets.

- Building and training machine learning models:

- Naive Bayes Classifier: Using CountVectorizer and TF-IDF Transformer.

- Linear Support Vector Machine (SVM): Using CountVectorizer and TF-IDF Transformer with SGDClassifier.

4. Model Evaluation:

- Evaluating models using accuracy scores and classification reports.

5. User Interaction:

- Allowing users to input headlines and predicting the news category using the trained SVM model.

**Results:**

- Achieved an accuracy of approximately **83%** on the test set using the Linear Support Vector Machine (SVM) model.

- The project enables users to input headlines interactively and predicts the corresponding news category using the trained SVM model.

**Future Improvements:**

- Experimenting with different machine learning algorithms and deep learning models to improve classification accuracy.

- Enhancing user interaction by deploying the model as a web application or chatbot for seamless access and usage.

By providing insights into the categorization of Indian English news articles, this project facilitates better understanding and analysis of news content for various stakeholders.