

Case Study: Text Classification of BBC News Articles

This dataset contains news headlines categorized into several classes such as business, technology, entertainment

Data Fields

- Article Id – Article id unique given to the record
- Article – Text of the header and article
- Category – Category of the article (tech, business, sport, entertainment, politics)

Using the techniques learned in the classes, build a text classification model to predict the category of a news article.

Instructions

1. Load the Dataset

- Import necessary libraries.
- Load the dataset ([BBC News Train.csv](#)) and explore its structure.

2. Preprocess the Text Data

- Normalize the text data by applying lowercasing and punctuation removal.
- Apply tokenization and stopwords removal to the normalized text data.
- Apply stemming and lemmatization to the tokenized text data.
- Transform the processed text data into numerical vectors using `CountVectorizer`.
- Create N-gram models (unigrams, bigrams, trigrams) and transform the text data using these models.

3. Build and Evaluate the Model

- Split the data into training and testing sets.
- Train a classifier (e.g., Logistic Regression, Naive Bayes, etc.) on the training set.

- Evaluate the model on the testing set using metrics like accuracy, precision, recall, and F1-score.

4. **Present the Results**

- Summarize the findings and model performance in a \

Deliverables:

- Preprocessed dataset
- Code and documentation of the steps followed
- A report summarizing the findings and model performance

All the very best..... 😊