News Article Classification Using NLP

Your Name

May 20, 2025

Abstract

This project aims to classify news articles into categories such as *World*, *Sports*, *Business*, and *Science & Technology* using natural language processing and machine learning methods. We evaluated multiple models including Logistic Regression, Naive Bayes, Random Forest, XGBoost, and MLP, and explored topic modeling with LDA.

1. Data Overview

We used a dataset consisting of a training and testing split with labeled news articles. The label mapping was as follows:

- 1 → World
- 2 → Sports
- 3 → Business
- 4 → Science & Technology

2. Preprocessing

- Combined the `Title` and `Description` columns to form the `Summary`.
- Removed missing values and irrelevant symbols.
- Applied text normalization: lowercasing, replacing symbols (e.g., $, %), removing extra spaces.
- Tokenization, stopwords removal, and lemmatization using WordNetLemmatizer.

3. Exploratory Data Analysis

Word clouds were generated for each class to identify dominant terms and topics. Label distribution was visualized using bar plots.

4. Feature Extraction

- Used Bag-of-Words model for basic vectorization.
- TF-IDF was applied to reduce the influence of frequent but less informative words.

5. Models Used and Accuracy

1. **Logistic Regression**:
   - Validation Accuracy: **0.719**
2. **Naive Bayes (Multinomial)**:
   - Cross-validation Accuracy: **0.8768**
   - Test Accuracy: **0.868**
3. **Random Forest**:
   - Best CV Accuracy: **0.8138**
4. **XGBoost Classifier**:
   - Best Parameters: learning_rate=0.3, max_depth=5, n_estimators=100
   - Best CV Accuracy: **0.8529**
5. **MLPClassifier (Neural Network)**:
   - Training Accuracy: **0.8143**
   - Validation Accuracy: **0.801**
   - Test Accuracy: **0.798**

6. Topic Modeling (LDA)

We applied Latent Dirichlet Allocation on the lemmatized and tokenized summaries. Topics extracted were coherent and interpretable:

- Topic 0: `ap game season new night win team lead year`
- Topic 1: `president bush election minister prime iraq vote leader`
- Topic 2: `microsoft window darfur talk sudan city peace government`
- ...

These topics revealed thematic clusters aligned with the original labels.

7. Conclusion

- Naive Bayes and XGBoost performed best on this task.
- Text preprocessing and proper feature extraction had significant impact.
- Topic modeling helped in interpretability and understanding latent themes.