News Article Classification - Final Report

Presentation clip

1. Introduction

This project focuses on building a machine learning model to classify news articles into predefined categories such as sports, politics, and technology. The classification helps in content filtering, recommendation, and content analysis in real-time applications.

2. Dataset

The dataset contains news articles along with their respective categories. Each article is labeled with a topic such as 'sports', 'politics', or 'technology'.

3. Data Preprocessing

Text data was cleaned by removing punctuation, numbers, and stopwords, and converting all text to lowercase. Tokenization and lemmatization were applied to normalize the text. Null values were removed from the dataset.

4. Feature Extraction

TF-IDF vectorization was used to convert the cleaned text data into numerical feature vectors. A maximum of 5000 features were extracted for model training.

5. Exploratory Data Analysis (EDA)

EDA was performed to visualize the distribution of categories and the most frequent words using bar plots and word clouds.

6. Model Development

Three models were developed and trained: Logistic Regression, Multinomial Naive Bayes, and Support Vector Machine (SVM). Each model was evaluated using classification metrics like accuracy, precision, recall, and F1-score.

7. Model Evaluation

All models were tested on a separate test set. The classification report and confusion matrix were generated for each model to assess performance. SVM and Logistic Regression performed the best with high F1-scores.

8. Conclusion

The SVM model was found to be the most accurate and reliable for classifying the news articles. The model can generalize well to unseen data and can be used in real-world applications for content categorization.

9. Future Work

Improvements can include using deep learning methods such as LSTM or BERT, and integrating real-time article streaming for live classification.