Mohammed Numan Suriya

15TH\_October

**News Article Classification Project**

Contents

[Introduction 2](#_Toc195348589)

[Data Collection and Preprocessing 2](#_Toc195348590)

[Feature Extraction 2](#_Toc195348591)

[Model Development & Training 2](#_Toc195348592)

[Model Evaluation 2](#_Toc195348593)

[Best Model Selection 2](#_Toc195348594)

[Key Findings 3](#_Toc195348595)

[Future Improvements 3](#_Toc195348596)

[Conclusion 3](#_Toc195348597)

# Introduction

This project involved building a machine learning model to categorize news articles into topics such as Sports, Politics, and Technology. We applied natural language processing (NLP) methods for tasks like text preprocessing, extracting features, training the model, and evaluating its performance.

# Data Collection and Preprocessing

* Imported the dataset of news articles.
* Cleaned text data (removing stop words, punctuation, and lowercasing).
* Handled missing data.
* Prepared the text for feature extraction.

# Feature Extraction

* Used TF-IDF vectorization to convert text into numerical features.
* Performed Exploratory Data Analysis (EDA) to understand category distributions.

# Model Development & Training

* Built classification models, such as: Logistic Regression and Naïve Bayes
* Tuned the models hyperparameters to improve performance.
* Used cross-validation to ensure robust evaluation.

# Model Evaluation

* Evaluated models using:
  + Accuracy
  + F1-Score
  + Used weighted F1-score for comparison.

# Best Model Selection

* Compared the performance of different models.
* Logistic Regression achieved the highest accuracy and F1-score.
* Selected Logistic Regression as the final model.

# Key Findings

* The use of feature extraction techniques, such as TF-IDF, plays a crucial role in improving model performance
* News articles can be accurately classified using NLP methods.
* Automation in classification enhances the efficiency of content organization.

# Future Improvements

* Explore the use of advanced deep learning models, such as LSTMs and BERT, to achieve improved accuracy.
* Broaden the dataset by incorporating additional categories.
* Enhance the feature engineering process for optimal results.

# Conclusion

This project showcases the application of machine learning and NLP methods for text classification, streamlining the process of managing content effectively. Furthermore, optimize accuracy with advanced deep learning models (LSTMs, BERT), expand the dataset with new categories, and refine feature engineering techniques.