

NEWS HEADLINE CLASSIFICATION PROJECT

(CLASSIFICATION MODE)

(MACHINE LEARNING AND DEEP LEARNING MODEL)

By: Hari Santhosh C

1. INTRODUCTION

The News Headline Classification Project aims to classify short news headlines into four categories: World, Sports, Business, and Sci/Tech. Machine learning and deep learning models were developed, and the final system was deployed as a Streamlit web application to provide real-time predictions.

A. APPROACH

The project follows a structured approach including data cleaning, preprocessing, vectorization, model training, evaluation, and deployment.

1. Data Loading and Cleaning

- Loaded datasets using pandas.
- Removed missing values and duplicates.
- Cleaned text by removing punctuation, URLs, numbers, emojis.
- Applied tokenization, stopword removal, and lemmatization.

2. Exploratory Data Analysis (EDA)

- Word clouds for text visualization.
- Count plots to observe class distribution.
- Category-wise sample inspection.

3. Preprocessing

- TF-IDF Vectorization converts text into weighted numerical vectors.
- Label Encoding maps text categories into numeric classes.

B. MODEL DEVELOPMENT

Four different models were trained: Random Forest, Multinomial Naive Bayes, XGBoost, and LSTM. Their performances were compared using multiple metrics.

C. MODEL EVALUATION

Models were evaluated using Accuracy, Precision, Recall, F1-Score, and Confusion Matrix to determine the best-performing algorithm.

D. HYPERPARAMETER TUNING

RandomizedSearchCV was used to optimize Naive Bayes parameters such as alpha and fit_prior, improving accuracy and prediction quality.

E. BEST MODEL SELECTED

The Tuned Multinomial Naive Bayes model delivered the highest accuracy and was selected for final deployment due to its efficiency and speed.

F. STREAMLIT DEPLOYMENT

A user-friendly Streamlit app was developed allowing:

- Single headline prediction
- CSV batch prediction
- Downloadable output

Backend loads TF-IDF vectorizer and trained NB model for production use.

G. CONCLUSION & PROJECT IMPACT

The News Headline Classifier automates fast and accurate categorization, improving workflows for media houses and content creators.