News Category Classification Using Machine Learning

# 1. Introduction

The objective of this project is to build a machine learning model that can accurately classify news articles into predefined categories such as politics, sports, and technology. The dataset consists of 50,000 labeled news articles collected from a reliable source.

# 2. Data Collection and Preprocessing

The dataset contains multiple columns, among which the 'short\_description' column was selected as the primary text for classification, and 'category' as the target label. Text preprocessing included steps such as lowercasing, removing punctuation and special characters, stopwords removal, tokenization, and lemmatization.

# 3. Exploratory Data Analysis (EDA)

EDA was performed to understand the distribution of categories and the characteristics of the text data. Bar plots were created to visualize the frequency of each category, and histograms were used to examine the distribution of word counts in news descriptions.

# 4. Feature Extraction

TF-IDF (Term Frequency-Inverse Document Frequency) was used for feature extraction, limiting the vocabulary to 5000 most important terms. This method helped convert the cleaned text data into numerical format suitable for machine learning algorithms.

# 5. Model Development

Three classification models were developed: Logistic Regression, Naive Bayes, and Support Vector Machine (SVM). The dataset was split into training and testing sets in an 80:20 ratio. Basic hyperparameter tuning was performed and cross-validation was used to evaluate the model robustness.

# 6. Model Evaluation

The models were evaluated using accuracy, precision, recall, and F1-score. Below is the evaluation summary:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1-Score |
| Logistic Regression | 0.7353 | 0.7372 | 0.7353 | 0.7359 |
| Naive Bayes | 0.7216 | 0.7230 | 0.7216 | 0.7218 |
| SVM | 0.7361 | 0.7363 | 0.7361 | 0.7358 |

SVM slightly outperformed the others with the highest accuracy and F1-score, making it the best model for this task.

# 7. Final Model Selection

The Support Vector Machine (SVM) model was selected as the final model based on its superior performance across all evaluation metrics. It offered the best balance of accuracy, precision, recall, and F1-score.

# 8. Conclusion

This project successfully built a robust machine learning pipeline for multi-class news classification using TF-IDF features and traditional classifiers. The final SVM model achieved approximately 73.6% accuracy and is capable of predicting the category of unseen news articles with high reliability.

# 9. References

• data\_news Dataset (source)

• <https://scikit-learn.org/>

• <https://www.nltk.org/>

# 10. Video Link

<https://drive.google.com/file/d/1jgCIpO1hVvtkgpj0x8HY1fia2tdc22ed/view?usp=sharing>