**ELEVATE LABS - AIML**

INTERNSHIP PROJECT REPORT

*PROJECT TOPIC:* NEWS ARTICLE CLASSIFICATION(FAKE/REAL)

*AUTHOR:* DHARSHINI R

*DATE:*  September 08, 2025

**INTRODUCTION:**

In the era of information, digital platforms are flooded with articles that need to be organized systematically. News articles cover diverse topics ranging from global politics to entertainment, and without proper classification, users may struggle to access relevant information. The aim of this project is to design an AI-powered solution that automatically categorizes articles. The approach relies on Natural Language Processing (NLP) to clean, preprocess, and analyze text data effectively. Classification algorithms such as Naive Bayes, Logistic Regression, and Neural Networks are tested to determine the best performing model. The system plays a significant role in applications like news recommendation engines, media monitoring tools, and content management systems. The project demonstrates how Machine Learning and AI contribute to handling real-world data challenges and improving user engagement.

**ABSTRACT:**

This project focuses on the development of a machine learning model for the classification of online news articles. With the continuous growth of digital platforms, large volumes of articles are published every minute across the globe. Categorizing these articles into relevant sections such as politics, sports, business, technology, and entertainment has become an essential task for publishers and news aggregators. By leveraging Artificial Intelligence and Natural Language Processing techniques, text is analyzed and mapped to categories. The main objective is to improve content recommendation systems, streamline information retrieval, and support digital media platforms. This report explains the methodology, tools, and steps involved in building the project and highlights its practical applications.

**TOOLS USED:**

Python was the primary programming language used for building the system due to its rich ecosystem for data science. Libraries such as Pandas and NumPy were applied for efficient data handling and preprocessing. Natural Language Toolkit (NLTK) and SpaCy supported essential NLP tasks like tokenization, stemming, lemmatization, and stop word removal. For model building, Scikit-learn was employed as it provides multiple classification algorithms. Matplotlib and Seaborn were used for data visualization and analysis of results. Jupyter Notebook served as the development environment to test and refine the models interactively. Dataset storage and processing were managed using CSV files and structured formats. Additional preprocessing included text normalization techniques. Together, these tools formed a robust framework to build and test the classification model successfully within the internship timeline.

**STEPS INVOLVED IN BUILDING THE PROJECT:**

1. **Data Collection** – Gather labled news datasets from open sources such as Kaggle or news APIs.
2. **Data Preprocessing** – Clean the raw text by removing punctuation, stop words, URLs, and special characters.
3. **Tokenization & Lemmatization** – Break text into tokens and normalize words to their base form using NLP tools.
4. **Feature Extraction** – Convert text into numerical vectors using TF-IDF or word embeddings.
5. **Train/Test Split** – Divide the dataset into training and testing sets to evaluate performance.
6. **Model Training** – Apply algorithms such as Logistic Regression, Naive Bayes, SVM, or Neural Networks.
7. **Model Evaluation** – Measure accuracy, precision, recall, and F1-score using the test data.
8. **Hyperparameter Tuning** – Optimize model parameters to improve classification performance.
9. **Final Model Selection** – Choose the best-performing model based on evaluation metrics.
10. **Deployment** – Integrate the trained model into a simple application or dashboard for real-time use.

**CONCLUSION:**

The News Article Classification project highlights the role of Artificial Intelligence and Machine Learning in solving real-world text categorization problems. By implementing NLP techniques and classification models, the system achieves accurate categorization of large volumes of data. It not only helps in managing digital content but also enhances user experiences by enabling faster information retrieval. The use of Python libraries and machine learning algorithms simplified the development process while ensuring reliable results. This project also demonstrates the importance of model evaluation metrics in improving accuracy. In practical applications, the system can be integrated into news portals, aggregators, and content management systems. Overall, the internship experience at Elevate Labs provided valuable exposure to AI-driven solutions for real-time challenges.