**News Category Classification using NLP and Machine Learning**

**Abstract:-** This project aims to classify news articles into four categories: World, Sports, Business, and Sci/Tech using Natural Language Processing (NLP) and Machine Learning techniques. We used the AG News dataset, applied text preprocessing, feature extraction using TF-IDF, and trained a Logistic Regression model. The model achieved an accuracy of approximately 85%, demonstrating that machine learning can effectively categorize textual news data. Visualizations such as WordClouds and class distribution charts were included to analyze the dataset.

**Introduction:-** The volume of news published online has increased exponentially, making automatic text classification an essential task for organizing information. Manual classification is inefficient, so NLP combined with Machine Learning provides an automated solution.

**Project Objective:**

* Preprocess and clean the news dataset.
* Extract features using TF-IDF.
* Train a machine learning model to classify news into categories.
* Evaluate the model using accuracy and F1-score.

**Dataset Description:**

Source: AG News Dataset

Training Data: ~120,000 news samples

Columns:

label – Category code (1 = World, 2 = Sports, 3 = Business, 4 = Sci/Tech)

title – News headline

description – Short description of the news

Categories:

-World

-Sports

-Business

-Sci/Tech

**Methodology:**

The project follows these steps:

Data Loading: Read train.csv into a DataFrame.

Data Preprocessing: Combine title + description, clean text (remove stopwords, punctuation, lowercase).

EDA: Class distribution chart and WordCloud.

Feature Extraction: Use TF-IDF to convert text into numeric features.

Model Training: Train Logistic Regression classifier.

Evaluation: Measure accuracy, precision, recall, F1-score, and plot confusion matrix.

Workflow Diagram:

Dataset → Preprocessing → TF-IDF → Model Training → Evaluation → Visualization

**Data Preprocessing:**

**Steps:**

Convert text to lowercase.

Remove punctuation and numbers.

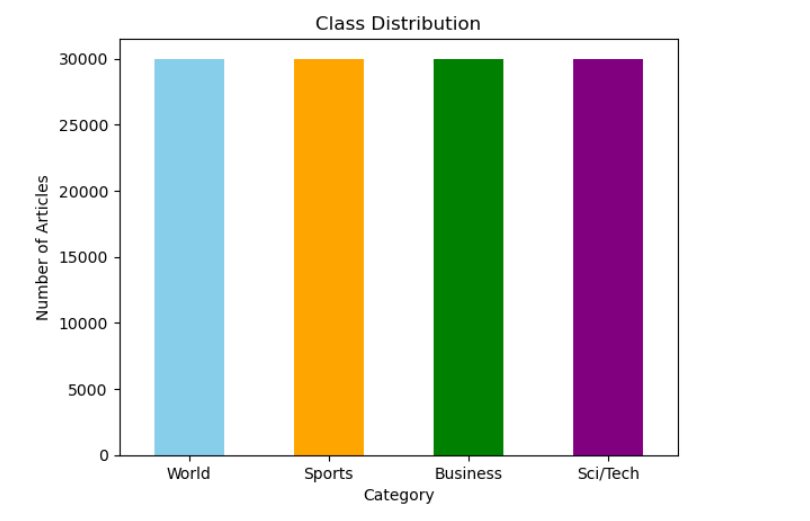
Remove English stopwords.

Combine title and description into one column.

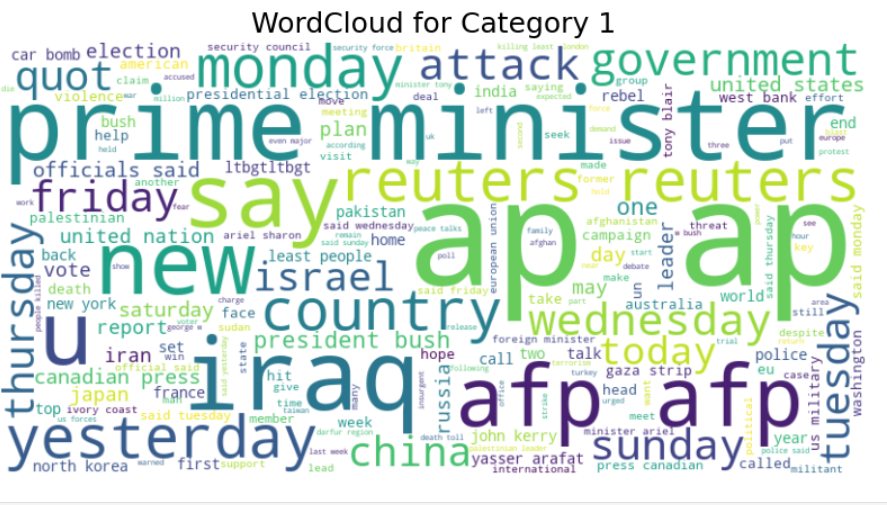
Apply TF-IDF vectorization (max\_features=5000).

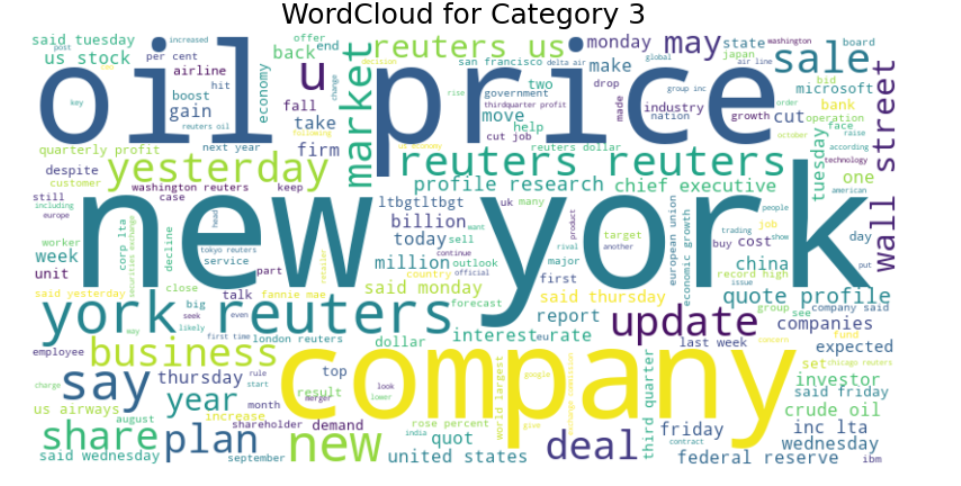
**Exploratory Data Analysis (EDA):**

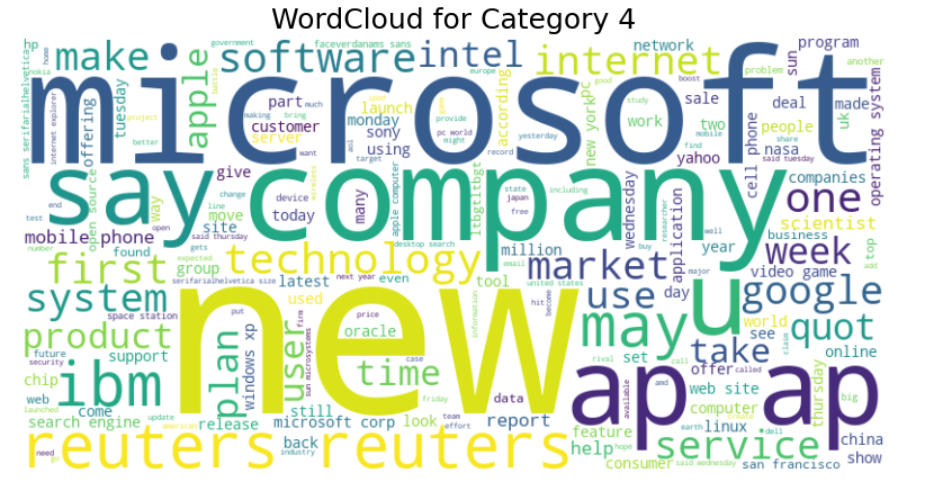
Class Distribution:



WordCloud:Highlights frequent words in news articles.







**Model Building:**

Algorithm: Logistic Regression

Training: 80% data for training, 20% for testing

Features: TF-IDF matrix

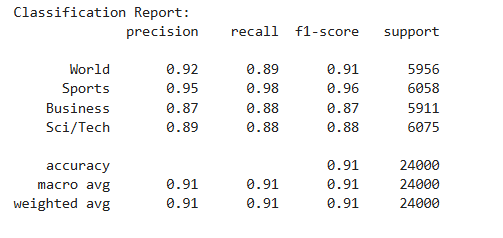
Target: label (numeric)

**Evaluation:**

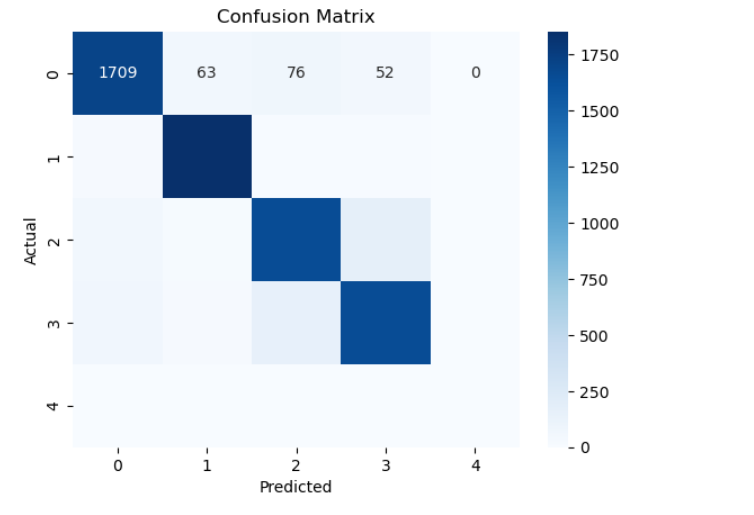
Metrics:

Accuracy: 90.67

Precision, Recall, F1-score:



Confusion Matrix:



**Results and Discussion:**

* The model correctly predicts most news categories with good accuracy.
* Sports category had the highest accuracy, while Business had slightly lower due to overlapping terms with World news.

### ****Future Enhancements:****

* Implement other models like **Naive Bayes, SVM** for comparison.
* Deploy the model using **Flask or Streamlit** for real-time predictions.
* Experiment with **deep learning models like BERT** for higher accuracy.

### ****Conclusion:**** This project demonstrates that NLP combined with machine learning is an effective method for classifying news articles. Logistic Regression with TF-IDF achieved good performance, making it suitable for real-world applications.