Course 7

# News Article Classification - Final Report

## NLP Project-Part-B

Satyajit Kumar
DATA SCIENCE

## Introduction:

In today's digital world, news articles are constantly being generated and shared across different platforms. For news organizations, social media platforms, and aggregators, classifying articles into specific categories such as sports, politics, and technology can help improve content management and recommendation systems. This project aims to develop a machine learning model that can classify news articles into predefined categories, such as sports, politics, and technology, based on their content.
Project Summary.

By automating this process, organizations can efficiently categorize large volumes of news articles, making it easier for readers to access relevant information based on their interests.

## Problem Statement:

The primary objective of this project is to build a classification model that can automatically categorize news articles into different predefined categories. The model will be trained using a labeled dataset of news articles and will output the most likely category (e.g., sports, politics, or technology) for any given article.

## 1.Data Collection and Exploration

- **Dataset**: 50,000 news articles evenly distributed across 10 categories
- **Categories**: BUSINESS, ENTERTAINMENT, FOOD & DRINK, PARENTING, POLITICS, SPORTS, STYLE & BEAUTY, TRAVEL, WELLNESS, WORLD NEWS
- **Key Features**:
    - Headline
    - Short description
    - Keywords
    - URL links

## 2. Data Preprocessing

1. **Text Combination**: Merged headline and short description into single text feature
2. **Text Cleaning**:
    - Converted to lowercase
    - Removed punctuation
    - Eliminated stopwords
3. **Feature Extraction**: TF-IDF vectorization (5,000 most frequent words)
4. **Train-Test Split**: 80-20 split with stratified sampling

## 3. Model Training

Implemented three classification models:

1. **Logistic Regression**
2. **Multinomial Naive Bayes**
3. **Linear Support Vector Machine (SVM)**

## 4. Model Evaluation Results

Performance Metrics Summary:

| Model | Accuracy | Precision (macro avg) | Recall (macro avg) | F1-Score (macro avg) |
|---|---|---|---|---|
| **Logistic Regression** | 0.7950 | 0.80 | 0.80 | 0.7953 |
| **Naive Bayes** | 0.7813 | 0.78 | 0.78 | 0.7815 |
| **Linear SVM** | 0.7891 | 0.79 | 0.79 | 0.7887 |

**Best Performing Categories Across All Models:**

- SPORTS (F1-score: 0.86-0.89)
- STYLE & BEAUTY (F1-score: 0.84-0.86)
- FOOD & DRINK (F1-score: 0.83-0.85)

**Key Findings**

1. **Model Performance**:
    - Logistic Regression achieved the highest overall accuracy (79.50%) and F1-score (0.7953)
    - Linear SVM showed slightly better performance than Naive Bayes
    - All models performed consistently well on sports and lifestyle categories
    - Wellness and parenting categories were most challenging across all models

**Conclusion**

All the three models performed well but the Logistic Regression emerged as the best-performing model among the three models that were tested.