# **CASE STUDY 4: NEWS ARTICLE CLASSIFICATION AND SUMMARIZATION**

**Prepared by:** URMISHIKHA DASH  
**Date:** 2/4/25

## **1. Executive Summary**

This report presents an automated system for **news article classification and summarization**. The objective is to efficiently categorize news articles into predefined topics such as **Politics, Sports, and Technology** while generating concise summaries for quick readability.

The system integrates **machine learning and deep learning models** to classify news articles and employs both **extractive and abstractive summarization techniques** to generate meaningful summaries.

### **Key Results:**

* **Preprocessing:** Tokenization, stopword removal, lemmatization.
* **Feature Extraction:** Word frequency analysis, TF-IDF.
* **Models Used:** Naïve Bayes, SVM, LSTM, CNN.
* **Best Performing Model:** **SVM (for classification), BART (for summarization).**

## **2. Objectives**

* **Automate article classification** into predefined categories.
* **Generate high-quality summaries** for enhanced readability.
* **Compare extractive and abstractive summarization** techniques.
* **Evaluate machine learning and deep learning models** for classification accuracy.

## **3. Methodology**

### **3.1 Data Collection and Preprocessing**

* **Dataset:** The **AG News Dataset** was selected due to its diverse news articles.
* **Cleaning Steps:**  
   Removed special characters, numbers, and URLs.  
   Converted text to lowercase for uniformity.  
   Tokenized text into words and lemmatized them.  
   Eliminated stopwords to retain meaningful words.

### **3.2 Data Analysis**

* **Category Distribution:** Analyzed to ensure balance in classification.
* **Text Structure Analysis:** Computed average word count and sentence length per category.
* **Word Frequency Visualization:** Used bar charts and word clouds to identify key themes.

### **3.3 Text Classification**

#### **Models Used:**

**Naïve Bayes:** A probabilistic classifier with moderate accuracy.  
  **Support Vector Machine (SVM):** Performed well by distinguishing categories with hyperplanes.  
  **Long Short-Term Memory (LSTM):** Captured contextual relationships in sequential data.  
  **Convolutional Neural Network (CNN):** Faster than LSTM but struggled with long-range dependencies.

#### **Evaluation Metrics:**

**Precision, Recall, and F1-score** were used to assess model effectiveness.  
  **SVM and LSTM delivered the best results for classification.**

### **3.4 Text Summarization**

#### **Techniques Applied:**

**Extractive Summarization:** Used **TF-IDF + PageRank** to extract key sentences.  
  **Abstractive Summarization:** Used **BART (Bidirectional and Auto-Regressive Transformers)** to generate summaries.

#### **Comparison and Performance Analysis:**

**Extractive Summarization:** Preserved original sentences but lacked flexibility.  
  **Abstractive Summarization:** More natural but occasionally introduced factual errors.

## **4. Results**

| **Task** | **Best Performing Model** | **Accuracy** |
| --- | --- | --- |
| **News Classification** | SVM, LSTM | **Highest Accuracy** |
| **Extractive Summarization** | TF-IDF + PageRank | **Preserved Key Sentences** |
| **Abstractive Summarization** | BART | **More Coherent Output** |

## **5. Discussion**

### **5.1 Strengths**

**Efficient news organization** through automated classification.  
  **Improves user experience** with concise, informative summaries.  
  **Deep learning models enhance accuracy** for both classification and summarization.

### **5.2 Limitations**

**Deep learning models require large datasets** and high computational power.  
  **Abstractive summarization may introduce inaccuracies** due to rephrasing errors.  
  **Extractive summarization lacks flexibility** in restructuring sentences.

### **5.3 Recommendations**

Optimize Transformer models for improved summarization accuracy.  
 Reduce computational costs for deep learning approaches.  
 Experiment with hybrid summarization (combining extractive and abstractive techniques).

## **6. Conclusion**

The system successfully **classified news articles** and generated **concise summaries**, improving news consumption efficiency. **SVM and LSTM models** performed best for classification, while **abstractive summarization (BART)**produced the most readable summaries.

Future improvements should focus on **refining Transformer models** and **reducing computational complexity** to enhance summarization accuracy and efficiency.

**Deliverables Submitted:**  
 Jupyter Notebook (cs4.ipynb)  
 Processed Dataset

## **Appendix A: Sample Data**

| **Article Snippet** | **Category** |
| --- | --- |
| "The government passed a new law..." | Politics |
| "The football team won the match..." | Sports |
| "New advancements in AI research..." | Technology |