AUTOMATED CLASSIFICATION OF NEWS ARTICLES BY CATEGORY REPORT

1. Introduction

The exponential growth of digital news content has made manual classification of articles inefficient. To address this, our project automates the classification of news articles into four categories: Politics, Technology, Entertainment, and Business. By leveraging machine learning (ML) and deep learning techniques, the project provides a scalable solution for content organization and user experience enhancement.

Our approach integrates traditional ML models (Logistic Regression, Random Forest, and SVM) and advanced deep learning models (LSTM and BERT). After a comprehensive evaluation, BERT was selected for its superior performance in understanding contextual meaning. The project also incorporates sentiment analysis and key phrase extraction, with predictions visualized using a Streamlit-powered dashboard.

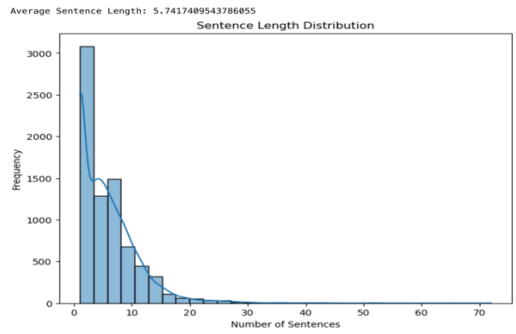
2. Methodology

2.1 Dataset Description

Dataset: News Category Dataset by Akash Gupta on Kaggle https://www.kaggle.com/datasets/akash14/news-category-dataset/data

The dataset consists of approximately 10,000 news articles. Each record includes:

- STORY: The content of the article (input text).
- SECTION: The target category (Politics, Technology, Entertainment, Business).



1. Sentence Length Distribution

The sentence length distribution graph shows that most articles contain 1 to 10 sentences, with the frequency decreasing for longer articles. The average sentence length is approximately 5.74 sentences, indicating that articles are generally short and concise. This skewed distribution reflects the brevity typical of news articles.

2. Category Distribution in Training Data

The bar chart highlights an imbalance in the dataset, where Category 1 has the most articles (around 2800), while other categories, especially Category 0, have fewer articles. Such imbalance can bias models toward the dominant category, requiring techniques like class weighting or data balancing to improve performance.

3. Word Cloud and Common Words
The word cloud reveals dominant words like "said", "year", "party", and "India", with a strong presence of political terms such as "election" and "congress". This suggests a significant proportion of the dataset is political news, alongside contributions from Business and Technology categories.

2.2 Preprocessing

The preprocessing stage involved multiple steps to prepare the text data for various machine learning and deep learning models. First, tokenization was performed to split the sentences into individual words. Next, stopword removal was applied to eliminate common words such as "the" and "and" that do not contribute significant meaning. Following this, lemmatization was used to reduce words to their base forms, ensuring consistency across variations. For traditional machine learning models, the cleaned text` was transformed into numerical features using the TF-IDF (Term Frequency-Inverse Document Frequency) vectorization technique. However, for BERT, minimal preprocessing was applied, including whitespace normalization and lowercasing, as BERT's tokenizer effectively handles raw text, punctuation, and word substructures natively. This dual preprocessing strategy ensured compatibility across all implemented models.

2.3 Models

Logistic Regression was selected as the baseline model for its simplicity and interpretability. Despite being a straightforward algorithm, it performed exceptionally well, with minimal misclassifications across all categories.

The Random Forest model, an ensemble-based approach, was employed to improve robustness and reduce overfitting. It proved reliable for text classification, with most misclassifications occurring within closely related categories.

The Support Vector Machine (SVM) model, trained using a linear kernel, performed similarly to Logistic Regression. It effectively differentiated subtle differences in text, particularly within categories like Politics and Business.

The Long Short-Term Memory (LSTM) model, designed for sequential data, captured long-term dependencies in text. Its performance steadily improved during training but required significantly more training time compared to traditional ML models.

Finally, the BERT (Bidirectional Encoder Representations from Transformers) model was fine-tuned for this task and outperformed all other models. BERT's ability to capture contextual relationships and nuances in the text made it the best-performing model, delivering state-of-the-art results.

Each model had its strengths, with BERT leading in accuracy and consistency, followed closely by SVM and Logistic Regression.

2.4 Sentiment Analysis

In addition to text classification, sentiment analysis was implemented to determine the emotional tone of news articles. Using NLTK's Sentiment Intensity Analyzer, sentiment scores were calculated for positive, negative, neutral, and compound values. The compound score, which ranges between -1 and +1, was used for classification. Articles were labeled as Positive (compound > 0.05), Neutral (between -0.05 and 0.05), or Negative (compound < -0.05). This approach provided insights into the overall emotional tone of the content.

Key phrase extraction was implemented using the RAKE (Rapid Automatic Keyword Extraction) algorithm to identify important phrases in the news articles. The method extracts top-ranked phrases based on word frequency and co-occurrence patterns, highlighting key contextual elements. This process enhances the understanding of article content by surfacing meaningful terms and themes.

3. Results and Evaluation

The performance of all models was evaluated using standard metrics: accuracy, precision, recall, and F1-score. Below is a summary of the results:

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	97	97	97	97
Random Forest	95	95	95	95
SVM	97	97	97	97
LSTM	94	94	94	94
BERT	97.9	97.91	97.9	97.9

Each model demonstrated its strengths, with BERT leading in overall accuracy and consistency, followed closely by SVM and Logistic Regression.

4. Conclusions

- Developed a robust system for automated classification of news articles into four categories: Politics, Technology, Entertainment, and Business.
- Leveraged BERT for superior performance, with sentiment analysis enhancing interpretability and user engagement.
- This project enhanced my understanding of text preprocessing

5. Future Research

- Expand categories for more granular classification of news articles.
- Enable users to upload files or process multiple articles in one go for increased usability.
- Add multilingual support to extend the system's reach for non-English news articles.
- Deploy the system as a fully functional web application for real-time article classification and interaction.